UNIVERSIDADE FEDERAL DO AMAZONAS
FACULDADE DE TECNOLOGIA
PROGRAMA DE PÓS-GRADUAÇÃO EM ENGENHARIA ELÉTRICA

JONILSON ROQUE DOS SANTOS

RECONHECIMENTO DAS CONFIGURAÇÕES DE MÃO DE LIBRAS BASEADO NA ANÁLISE DE DISCRIMINANTE DE FISHER BIDIMENSIONAL UTILIZANDO IMAGENS DE PROFUNDIDADE.

MANAUS
2015

UNIVERSIDADE FEDERAL DO AMAZONAS
FACULDADE DE TECNOLOGIA
PROGRAMA DE PÓS-GRADUAÇÃO EM ENGENHARIA ELÉTRICA

JONILSON ROQUE DOS SANTOS

RECONHECIMENTO DAS CONFIGURAÇÕES DE MÃO DE LIBRAS BASEADO NA ANÁLISE DE DISCRIMINANTE DE FISHER BIDIMENSIONAL UTILIZANDO IMAGENS DE PROFUNDIDADE.

Dissertação apresentada ao Programa de Pós-Graduação em Engenharia Elétrica da Universidade Federal do Amazonas, como requisito parcial para obtenção do título de Mestre em Engenharia Elétrica na área de concentração Controle e Automação de Sistemas.

Orientador: Prof. Dr. Cícero Ferreira Fernandes Costa Filho
Co-Orientadora: Profª. Drª. Marly Guimarães Fernandes Costa

MANAUS
2015

Ficha Catalográfica

Ficha catalográfica elaborada automaticamente de acordo com os dados fornecidos pelo(a) autor(a).

Santos, Jonilson Roque dos
S237r Reconhecimento das configurações de mão de LIBRAS baseado na Análise de Discriminante de Fisher bidimensional utilizando imagens de profundidade. / Jonilson Roque dos Santos. 2015
95 f.: il. color; 31 cm.

Orientador: Cícero Ferreira Fernandes Costa Filho
Coorientador: Marly Guimarães Fernandes Costa
Dissertação (Mestrado em Engenharia Elétrica) - Universidade Federal do Amazonas.

1. Língua Brasileira de Sinais. 2. Reconhecimento de padrões. 3. Kinect®. 4. 2D2LDA. 5. , k-vizinhos mais próximos (k-NN). I. Costa Filho, Cícero Ferreira Fernandes II. Universidade Federal do Amazonas III. Título

JONILSON ROQUE DOS SANTOS

RECONHECIMENTO DAS CONFIGURAÇÕES DE MÃO DE LIBRAS BASEADO NA ANÁLISE DE DICRIMINANTE DE FISHER BIDIMENSIONAL UTILIZANDO IMAGENS DE PROFUNDIDADE.

Dissertação apresentada ao Programa de Pós-Graduação em Engenharia Elétrica da Universidade Federal do Amazonas, como requisito parcial para obtenção do título de Mestre em Engenharia Elétrica na área de concentração Controle e Automação de Sistemas.

Aprovado em 04 de setembro de 2015.

BANCA EXAMINADORA

Cícero Ferreira Fernandes Costa Filho, Presidente

Universidade Federal do Amazonas-UFAM

Prof. Dr. Cícero Augusto Mota Cavalcante, Membro

Universidade Federal do Amazonas-UFAM

Prof. Dr. Jozias Parente de Oliveira, Membro

Universidade do Estado do Amazonas-UEA

AGRADECIMENTOS

Primeiramente, gostaria de agradecer à minha família pelo apoio e compreensão, a meus pais pela formação dada e por acreditarem nessa nova conquista.

Agradeço em especial aos meus orientadores, Prof. Dr. Cícero Ferreira Fernandes Costa Filho e Prof.ª Dra. Marly Guimarães Fernandes Costa, pela paciência, dedicação e, principalmente, pela confiança no meu trabalho e capacidade.

Aos professores, funcionários do Centro de Pesquisa e Desenvolvimento de Tecnologia Eletrônica e da Informação – CETELI pela disponibilização de toda a estrutura física, laboratorial e dos materiais de pesquisa para o desenvolvimento deste trabalho.

Parte dos resultados apresentados neste trabalho foram obtidos através do Projeto de Pesquisa e formação de recursos humanos, em nível de graduação e pós-graduação, nas áreas de automação industrial, softwares para dispositivos móveis e TV Digital, financiado pela Samsung Eletrônica da Amazônia Ltda., no âmbito da Lei no. 8.387 (art. 2º) /91.

À Direção da Escola Estadual de Educação Especial Augusto do Santos Carneiro e a todos os alunos voluntários pela prestimosa ajuda a esse trabalho.

Ao professor Marcos Roberto, pelas discussões iniciais e aula de LIBRAS.

Por fim agradeço aos amigos, em especial Endrews Souza, Robson Souza e Márcio Rodrigues que acompanharam minha jornada no decorrer de meus estudos e dedicação ao mestrado.

RESUMO

As pessoas surdas comunicam-se com outras pessoas por meio da Língua de Sinais. Essa interação restringe-se somente a pessoas que conhecem a Língua que, via de regra, são as pessoas surdas. O fato é que existem muitas pessoas, notadamente das áreas de saúde, educação e lazer, que necessitam interagir com os surdos usando Língua de Sinais e possuem pouca ou nenhuma proficiência na Língua de Sinais. Então, a inclusão social do surdo é seriamente afetada, pois ele não é capaz de se fazer entender. Esta dissertação apresenta uma metodologia para o reconhecimento automatizado dos gestos que representam as configurações de mãos da Língua Brasileira de Sinais - LIBRAS. A abordagem inicial consistiu na construção conjunta de um banco de imagens das configurações de mão capturada através de uma câmara de profundidade, Kinect®. A região de interesse, a mão realizando o gesto, foi extraída utilizando-se as seguintes técnicas: *K-means* e Transformada de Distância. O processo de Reconhecimento dos gestos foi dividido em duas etapas: extração de características e classificação dos gestos. Dessa forma, foi aplicado a técnica de redução de dimensionalidade, 2D2LDA para a obtenção de um conjunto de características, as quais foram submetidas a um classificador, o k-vizinhos mais próximos (kNN). O método proposto é capaz de segmentar e reconhecer as 61 configurações de mão da Língua Brasileira de Sinais. A taxa média de acerto alcançada foi de 96,10%. Como o dispositivo de captura é insensível a luminosidade, fundo e cores das roupas e da pele, a aplicação desenvolvida adapta-se sem necessidade de modificações a qualquer outro ambiente de captura.

Palavras chaves: Língua Brasileira de Sinais; Reconhecimento de padrões; Kinect®; 2D2LDA, *K-means, k-vizinhos mais próximos (k-NN).*

ABSTRACT

Deaf people communicate with other people using sign language. This communication is limited to people with knowledge in the language that, usually, are other deaf people. The fact is that there are too many people interacting with deaf people in education, health and leisure areas that are not proficient in sign language. Then, the inclusion of deaf people is seriously affected, because they are unable to make themselves understood. This study presents a methodology for automatic gesture recognition which represents hands settings from Brazilian Language of Signs - LIBRAS. The first approach consisted in a constructing of hands settings image database captured by depth camera, Kinect®. The region of interest, hands making gesture, was extracted using the following techniques: K-means and Distance Transformation. The recognition part was divided in two steps: feature extraction and gesture classification. This way, the dimensionality reduction technique was applied, 2D2LDA to obtain a features set, which was submitted to a classifier, k-nearest neighbor. The proposed system is able to segment image and recognize whole 61 settings of Sign Language. The average hit rate achieved was 96.10%. As the capture device is insensitive to light, background and colors of clothes and skin, the developed application adapts without modifications to any other capture environment.

Keywords: Brazilian Sign Language; Recognition Pattern; Kinect®; 2D2LDA; K-means; k-nearest neighbor (kNN).

LISTA DE ILUSTRAÇÕES

LISTA DE QUADROS

LISTAS DE TABELAS

SUMÁRIO

1 INTRODUÇÃO

A comunicação entre as pessoas é efetuada de diversas formas: oral, escrita ou gestual. A língua de sinais é uma forma gestual que possibilita as pessoas surdas interagirem com outras pessoas. Os surdos, no entanto, restringem a sua comunicação apenas a pessoas que conhecem a língua de sinais. O fato é que existem muitas pessoas, notadamente das áreas de saúde, educação e lazer, que necessitam interagir com os surdos usando Língua de Sinais e possuem pouca ou nenhuma proficiência na Língua de Sinais. Então a inclusão social do surdo é seriamente afetada, pois ele não é capaz de se fazer entender. Para atender a demanda de comunicação dos surdos com a sociedade, tem sido criados sistemas automáticos que têm o intuito de reconhecer os gestos da língua de sinais (MAUNG, 2009; CARNEIRO, 2010). A maioria desses sistemas é baseada em câmeras sensíveis a luz visível e, portanto, sensíveis às condições de luminosidade, *background* e tom de pele do usuário, o que acarreta muitos erros na segmentação da região de interesse. Com o intuito de minimizar a dependência com a luminosidade, diversas abordagens foram adotadas, como a utilização de luvas coloridas ou de luvas sensórias (MARAQA *et al.*, 2012), que facilitam a identificação da orientação e posição da mão no espaço, mas que exigem a interligação da luva a um computador. De acordo com Mitra e Acharya (2007), o uso desses aparatos diminui a naturalidade da interação.

Atualmente, com a criação de novas tecnologias de baixo custo, como a câmara Kinect® da Microsoft®, que fornece além da imagem *true color*, uma mapa de profundidade da cena, consegue-se trabalhar com baixos níveis de luminosidade e ter êxito na segmentação do gesto (SHOTTON *et al.*, 2013).

No Brasil, a Língua de sinais oficial utilizada pelos surdos é a Língua Brasileira de Sinais (LIBRAS). De acordo com Soares (2005), a Língua Brasileira de Sinais é um sistema

linguístico legítimo e natural, utilizado pela comunidade surda brasileira, de modalidade gestual-visual e com estrutura gramatical independente da Língua Portuguesa falada no Brasil.

Ela é derivada tanto de uma língua de sinais autóctone, quanto da Língua gestual francesa; por isso, é semelhante a outras línguas de sinais da Europa e da América. De acordo com Goldfeald (2003), a LIBRAS não é a simples gestualização da língua portuguesa, e sim uma língua à parte. Essa afirmativa é comprovada pelo fato de que em Portugal usa-se uma língua de sinais diferente, a Língua Gestual Portuguesa (LGP).

A LIBRAS é dotada de toda a complexidade e utilidade encontrada nas línguas orais e, assim como elas, possui gramática própria, com regras específicas em seus níveis linguísticos, fonológico, morfológico e sintático (MARCOTTI *et al.*, 2007). Entretanto, somente a partir de 24 de abril de 2002 a Libras foi reconhecida como meio legal de comunicação e expressão da comunidade surda do Brasil, de acordo com a lei Nº. 10.436, decretada pelo Congresso Nacional e sancionada pelo presidente da república.

A estrutura fonológica da LIBRAS, isto é a maneira como são formadas as palavras, possui cinco parâmetros. As combinações dos mesmos formam as palavras da Língua. Segundo Guimaraes *et al*. (2010), a Língua Brasileira de Sinais é constituída fonologicamente pelos seguintes parâmetros globais: o ponto de articulação, a configuração da mão, o movimento da mão, a orientação da palma da mão e as expressões faciais e corporais.

Muitos pesquisadores trabalham no reconhecimento de um conjunto finito de palavras, como, Chao *et al*., (2013), ou no reconhecimento do alfabeto de LIBRAS, como (Bragatto *et al.,* (2006). A primeira abordagem é restrita, pois é necessário um conjunto de treinamento muito grande para treinar, de forma robusta, todos os sinais da Língua. Na segunda abordagem, Sánchez (1989 apud KUSASKI e MORAES, 2009, p. 3417) diz que “os surdos, de forma diferente dos ouvintes, não podem aprender o som das letras porque não ouvem e não podem

fazer uso do mecanismo alfabético para extrair significado do escrito e comunicar-se com outras pessoas".

Este trabalho apresenta uma contribuição para o reconhecimento automatizado da LIBRAS, centrando inicialmente a sua atenção no reconhecimento de um dos parâmetros globais da LIBRAS, a configuração da mão (CM), como parte integrante do problema principal, que é o reconhecimento da Língua de Sinais. A principal vantagem da proposta em comparação com outros sistemas é a capacidade do reconhecimento de configurações de mãos complexas utilizando a Análise de Discriminante de Fisher bidimensional para geração de características que podem diferenciar os gestos entre si.

A Língua Brasileira de Sinais possui 61 possíveis CM, as quais são apresentadas na Figura 1.

Figura 1- Configurações de mão da LIBRAS. (PIMENTA e DE QUADROS, 2010).

1.1 OBJETIVO GERAL

Propor e implementar um método para reconhecimento das 61 configurações de mãos da LIBRAS, utilizando informações de profundidade obtidas através do sensor Kinect®.

1.2 OBJETIVOS ESPECÍFICOS

1. Adquirir uma base de dados das configurações de mão da LIBRAS com o sensor Kinect®;
2. Propor e implementar um método para segmentação das 61 CM da LIBRAS;
3. Caracterizar a utilização da extensão bidirecional da Análise Discriminante de Fisher, 2D2LDA, para extração de características significativas das CM da LIBRAS, visando o reconhecimento das mesmas;
4. Avaliar o desempenho da técnica de reconhecimento de padrões *k–vizinhos mais próximos* (kNN - *k-Nearest Neighbor*), no problema estudado, utilizando mais de uma medida de distância, para o reconhecimento das CMs da LIBRAS.

1.3 ORGANIZAÇÃO DO TRABALHO

A organização dessa dissertação segue as divisões a seguir:

- Introdução (Seção 1);
- Revisão Bibliográfica (Seção 2);
- Referencial Teórico (Seção 3);
- Materiais e Métodos (Seção 4);
- Resultados e Discussão (Seção 5);

- Conclusões e Trabalhos futuros (Seção 6);
- Referências;
- Apêndice: trabalho publicado.

Na seção 2 são apresentados os artigos científicos sobre reconhecimentos de gesto usando visão computacional com câmeras convencionais, luvas coloridas e sensores de profundidade que foram revisados.

A seção 3 aborda aspectos teóricos de técnicas envolvidas com o processo de reconhecimento de gestos. Por exemplo: é analisada a utilização de um algoritmo não supervisionado na segmentação de um objeto de interesse; é descrita uma técnica capaz de extrair características de forma a maximizar a separabilidade das classes e descreve-se o mecanismo do algoritmo de classificação usado neste trabalho.

A seção 4 relata os materiais e métodos usados. A seção inicia descrevendo as principais características do banco de dados adquirido, como quantidade de imagens e resolução das imagens adquiridas. Em seguida, descreve-se as principais etapas do método de reconhecimento de padrões utilizado para o reconhecimento das CM de LIBRAS: a etapa de segmentação dos gestos, a padronização das imagens, a técnica de seleção de características utilizada, 2D2LDA, e o método de reconhecimento *k–vizinhos mais próximos*.

A seção 5 apresenta os resultados obtidos através do método de reconhecimento de padrões proposto na seção anterior e discute os resultados obtidos.

Na seção 6 é analisado se os objetivos da pesquisa foram alcançados e sugere trabalhos futuros para o tema em estudo.

2 REVISÃO BIBLIOGRÁFICA

A utilização de técnicas de visão computacional na tarefa de reconhecimento de gestos de Línguas de sinais está atraindo cada vez mais a atenção dos pesquisadores. Os estudos publicados modelam a Língua gestual de várias maneiras. Alguns autores estudam a Língua por completo (RAKUN *et al.*,(2013); CHAO *et al.*,(2013)), ou seja, modelam toda estrutura fonológica, enquanto outros restringe-se apenas ao reconhecimento do alfabeto da Língua de sinais (BRAGATTO, RUAS E CASTRO (2006); CARNEIRO (2010); MARAQA *et al*. (2012)) . O reconhecimento de gestos com as mãos é uma tarefa extremamente difícil devido à complexidade da estrutura da mão humana e dos movimentos. Sensores tradicionais são sensíveis às condições do meio no qual a cena está sendo adquirida, como fundo, iluminação e tom de pele. Para lidar com as dificuldades referidas, várias pesquisas propuseram o uso de luvas coloridas, luvas sensoriais, de mais de uma câmera, e, mais recentemente, passaram a explorar a informação 3D adquirida por sensores estéreos passivos (RAKUN *et al.*,(2013); CHAO *et al.*,(2013); (PORFIRIO *et al.*(2013);ZHOU *et al.*(2013)).

Esta seção visa discorrer sobre as principais abordagens utilizadas na literatura para o reconhecimento de Língua de sinais com o uso de câmeras de intensidades, bem como sobre estudos que utilizam luvas coloridas. Por fim serão abordados sistemas que adotam o uso de sensores de profundidade com o intuito de reconhecimentos das mais diversas línguas de sinais existentes no mundo.

As seções dividem-se em: seção 2.1, que apresenta o reconhecimento de língua de sinais e fazem uso de câmeras convencionais; seção 2.2, que apresenta abordagens onde o reconhecimento de gestos é facilitado pelo uso de luvas; seção 2.3, que apresenta trabalhos com o uso de sensor de profundidade. O Quadro 2, ao final desta seção, apresenta um resumo da revisão bibliográfica de onze artigos sobre o reconhecimento de gestos.

2.1 RECONHECIMENTO DE LÍNGUA DE SINAIS COM O USO DE CÂMERAS CONVENCIONAIS.

O uso de câmera convencional como forma de aquisição de imagens foi empregado em pesquisas pioneiras na área de reconhecimento de gestos. Várias foram as técnicas empregadas nesses desenvolvimentos.

Ribeiro (2006) desenvolveu um sistema de reconhecimento em tempo real de 10 gestos não associados à Língua de Sinais. A aquisição foi realizada a partir de uma sequência de imagens de vídeo, obtida com uma câmera (*webcam)* em ambiente interno e fonte de luz artificial. No pré-processamento os autores utilizam um filtro gaussiano com o objetivo de suavizar as imagens. Para segmentação da mão, o estudo usou processos de eliminação do *background* associado com a técnica de misturas de gaussianas e limiarização simples. Para correção de erros da segmentação, o autor utiliza filtros morfológicos. No término da segmentação, o autor emprega técnicas de detecção do contorno e remoção do antebraço através da Transformada de Distância Euclidiana (TDE). O processo de classificação dos gestos foi dividido em duas etapas: extração de características e classificação propriamente dita. As características extraídas incluíram os sete momentos de HU e sete propriedades geométricas do contorno da mão. A etapa de classificação foi feita através de casamento de padrões *(template matching)* utilizando a medida euclidiana para calcular a distância de um novo padrão em relação a um modelo. O sistema desenvolvido obteve taxa de acerto global de 94,34%.

Maung (2009) e Carneiro (2010) utilizam Redes Neurais Artificiais (RNA) para reconhecimento de gesto do alfabeto da Língua de sinais de Myanmar e de LIBRAS, respectivamente. O primeiro estudo citado utilizou a técnica de histograma de orientação local para definir as entradas da rede neural. A taxa média de reconhecimento obtida foi de 90%.

No segundo estudo citado, construiu-se um banco de dados com 7800 imagens no padrão RGB (*Red, Green, Blue*) das 26 letras do alfabeto LIBRAS. Essas imagens foram obtidas com a colaboração de seis pessoas, num ambiente com iluminação e fundo controlados, o que simplifica sobremaneira o processo de segmentação da mão. Na etapa de pré-processamento, os autores fazem uso do algoritmo de especificação de histograma. A etapa de segmentação da mão é efetuada através de limiarização simples nos espaços de cor YCbCr (Luminância, Crominância Azul, Crominância Vermelha), RGB e HSV (*Hue*, *Saturation*, *Value*) e por um modelo de probabilidade que é gerado através da componente Cb e Cr da imagem. Na etapa de pós-processamento, com o objetivo de filtrar erros da segmentação, aplica-se o algoritmo de crescimento de região. A classificação dos gestos foi feita a partir de duas etapas: extração de características e classificação. Na extração de características usou-se dois tipos de atributos: momentos invariantes de Hu e Descritores de *Fourier*. Na segunda etapa o autor realiza, inicialmente, uma pré-classificação através de uma técnica de agrupamento, de tal forma que cada conjunto tenha um número reduzido de classes. Com esse objetivo, utilizou-se a rede *Self-Organized-Maps* (SOM). A Classificação final é feita direcionando cada agrupamento para uma rede neural artificial supervisionada. O melhor resultado obtido foi uma taxa de acerto de 82,67%. Para a obtenção desse resultado utilizou-se 50 Descritores de *Fourier* como variáveis de entrada do classificador.

2.2 RECONHECIMENTO DE GESTO COM LUVAS SENSORIAS E LUVAS COLORIDAS.

Em sistemas baseados em luvas sensoriais, as informações utilizadas pelo classificador são provenientes de sensores localizados em diversas partes da mão. A partir das informações dos sensores, importantes informações utilizadas pelo classificador, como posição, velocidade

e orientação da mão, são extraídas. Já as luvas de tecido, que podem ser totalmente ou parcialmente coloridas, têm como principal objetivo facilitar a segmentação do objeto de interesse.

Bragatto, Ruas e Castro (2006), fazem o reconhecimento de 26 letras do alfabeto LIBRAS. Para tanto, utilizam uma luva com seis cores para diferenciar os dedos da palma da mão, em imagens adquiridas com fundo complexo e sem controle de luminosidade. Na etapa de segmentação, treina-se um classificador RNA Perceptron Multicamadas (MLP) para classificação dos *pixels* das imagens. Filtros morfológicos são empregados para remoção de ruídos. Para classificação dos gestos são extraídas quatro características de cada dedo, provenientes da análise de componentes principais morfológicas descritas em (LAMAR *et al.*, 1999). Diversos experimentos foram realizados, sendo a melhor a taxa de reconhecimento obtida de 99,2%. O trabalho não descreve como modelou os sinais dinâmicos que apresenta a LIBRAS e nem o conjunto de dados que utilizou no experimento.

Maraqa *et al*. (2012) utilizaram três redes neurais recorrentes para reconhecer os gestos estáticos da Língua de Sinais Árabe. Duas pessoas realizaram 30 repetições de cada gesto usando uma luva colorida e câmera digital, resultando em um banco de dados com 1800 imagens. Para facilitar o processo de segmentação, os autores empregaram uma codificação para as cores de diversas regiões da mão (ver Figura 2). Existem seis camadas de cor, sendo uma para cada dedo e uma para o pulso. Para entrada no classificador, são extraídas trinta características que expressam ângulos de orientação e posição das pontas dos dedos em relação ao pulso (Figura 3). As imagens do estudo são coloridas e foram divididas em dois grupos: 900 imagens no conjunto de treinamento e 900 imagens no conjunto de teste. Na etapa de classificação foram utilizadas RNA recorrentes, RNA com arquitetura *Elman* e RNA com arquitetura Jordan. Os melhores resultados foram obtidos com RNA recorrentes, uma taxa de reconhecimento de 95,11%. Já com RNA com arquitetura *Elman*, obteve-se uma taxa de

reconhecimento de 89,67%, enquanto que com RNA com arquitetura *Jordan*, obteve-se uma taxa de reconhecimento de 84,54%.

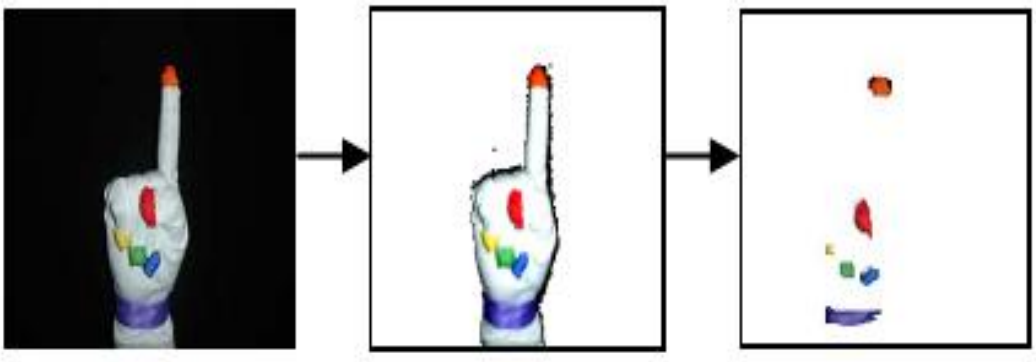


Figura 2-Modelo de cor da luva. (MARAQA *et al*., 2012)

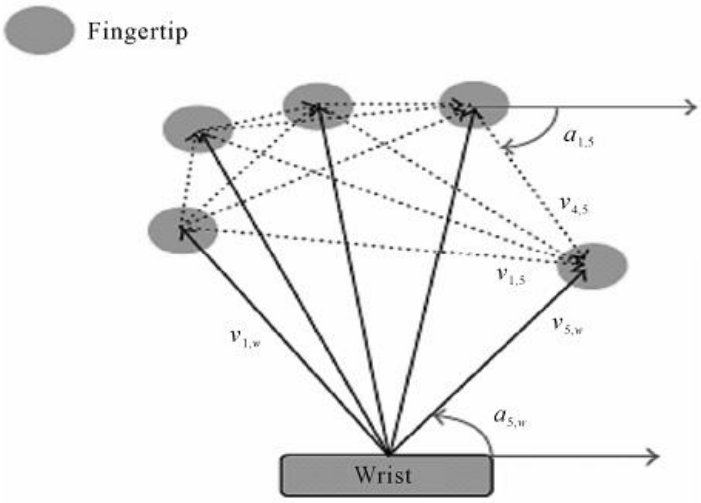


Figura 3- Posição e orientação de cada dedo da luva colorida. (MARAQA *et al*., 2012)

2.3 RECONHECIMENTO DE LÍNGUA DE SINAIS COM SENSOR DE PROFUNDIDADE.

Os métodos tradicionais estão longe de ser satisfatórios para aplicações na vida real. A qualidade da imagem capturada é sensível às condições de luz, fundos complexos e tons de pele. Às vezes, tais métodos requerem calibração o que prejudica a naturalidade de uma aplicação *Human Computer Interaction* (HCI).

O surgimento da tecnologia de mapeamento 3D com custo relativamente barato, deu um novo estímulo à pesquisa na área de rastreamento de objetos e reconhecimento de gestos. Um dos mais populares dispositivos usados para esses fins é o Kinect® da empresa Microsoft®. A seguir aborda-se o conteúdo de alguns trabalhos que empregaram o sensor de profundidade Kinect® e outros sensores com a mesma função.

Malassiotis e Strintzis (2008) trabalharam com reconhecimento de 20 posturas de mãos representando números (0 a 9) e letras da Língua de Sinais Alemã. Para constituição do conjunto de treinamento, três pessoas participaram da aquisição de imagens cor e de profundidade. Para cada pessoa e para cada postura foram adquiridas 50 imagens. O conjunto de teste foi separado em duas sessões. Em cada sessão a postura da mão foi diferente. Em cada sessão foram obtidas 2000 imagens, através da colaboração de mais duas pessoas. O autor realizou dois tipos de experimentos. No primeiro usou apenas a silhueta da mão 2D em dois tipos de imagens, RGB e profundidade. Nas imagens RGB aplicou segmentação por cor, conforme proposto por Yin e Xie (2007). Nas imagens de profundidade, para segmentação do braço, utilizou-se algoritmos hierárquicos. Nos dois tipos de imagens, o pós-processamento realizou a remoção do antebraço através do uso de misturas de duas gaussianas. Na fase de extração de características utilizou-se descritores elípticos de *Fourier* do contorno das imagens. O algoritmo de classificação empregado foi o k-vizinhos mais próximos. No Quadro 1 apresentam-se as taxas de reconhecimento obtidas para as abordagens 2D e 3D. Conforme pode ser visto, as taxas de reconhecimento 2D foram levemente superiores às taxas de reconhecimento 3D. Acredita-se que esse resultado seja devido a qualidade da máscara da mão utilizada.

Quadro 1- Taxa de reconhecimento de gestos correspondente a letras do alfabeto Alemão, usando descritores elípticos de *Fourier* 2-D.

Números de descritores elípticos de *Fourier*	Primeira sessão		Segunda sessão	
	RGB (%)	Profundidade (%)	RGB (%)	Profundidade (%)
10	77	76	72	67
15	79	78	73	71
20	83	81	80	74

Fonte: Malassiotis e Strintzis (2008).

Rakun *et al.* (2013) propuseram um sistema de reconhecimento de 10 palavras da Língua Indonésia (SIBI), que emprega quatro componentes fonológicas para diferenciar os sinais. O banco de dados é formado por 12 amostras de cada sinal. Os autores não propõem nenhum método para segmentação, pois utilizam funções de rastreio do kit de desenvolvimento de *software* (SDK) do Kinect® para detectar a mão. A partir da imagem de profundidade são extraídas as seguintes características do gesto: área, centroide, eixo maior e menor, orientação e menor polígono convexo. A partir da informação de cor da imagem é extraída a Transformada Discreta Cosseno com Correlação cruzada (DCTCC). A partir do rastreio das articulações do esqueleto do usuário com SDK calcula-se os ângulos de quatro delas (cotovelo direito, cotovelo esquerdo, mão direita e mão esquerda) em relação ao ponto central do ombro conforme ilustra a Figura 4. Os algoritmos de classificação usados foram *Generalized Learning Vector Quantization* (GLVQ) e *Random Forest*. Diversos experimentos com várias combinações entres as características extraídas foram realizados. A melhor taxa de reconhecimento média foi de 94,37% obtida com o classificador *Random Forest* associando as características de propriedades geométricas da região com os ângulos das articulações. Embora a abordagem dos autores tenha visado atender as quatro componentes fonológicas do SIBI, eles se limitam a apenas 10 sinais.

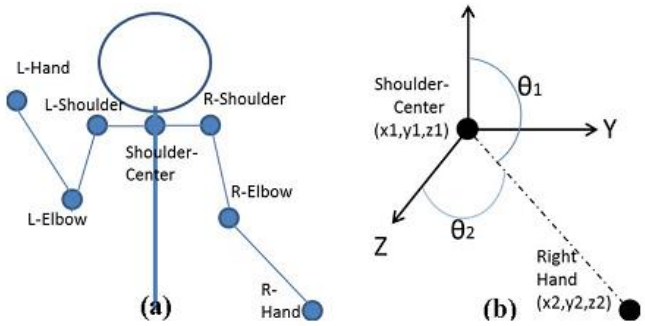


Figura 4- Representação de um esqueleto pelo Kinect® (a) articulações rastreadas pelo Kinect® (b) ilustração da posição das articulações com relação ao centro do ombro. (RAKUN *et al.*, 2013).

Chao *et al.* (2013) construíram uma base de dados com 73 sinais da Língua Americana de Sinais (ASL). Nove pessoas pousaram para câmera, com cada um repetindo todos os sinais três vezes. Foram coletadas imagens RBG, mapas de profundidade e informações de posicionamento das articulações do corpo, conforme mostrado na Figura 5. Diversas características foram extraídas, como: Histograma de Gradiente Orientado (HOG) e informações do Kinect® como pose do corpo, forma da mão e movimentação da mão. Na etapa de classificação utilizou-se máquinas de vetores de suporte latente, uma vez que esse método é capaz de encontrar quadros discriminativos e representativos de sinais em cada conjunto de vídeo. Os autores reportam uma acurácia média de 82,3%, com características extraídas de HOG e, 86,0% de acurácia quando se utilizou características extraídas do HOG associadas a características obtidas a partir do sensor Kinect®. Os resultados mostram que o Kinect® pode efetivamente contribuir na classificação de Línguas de Sinais.

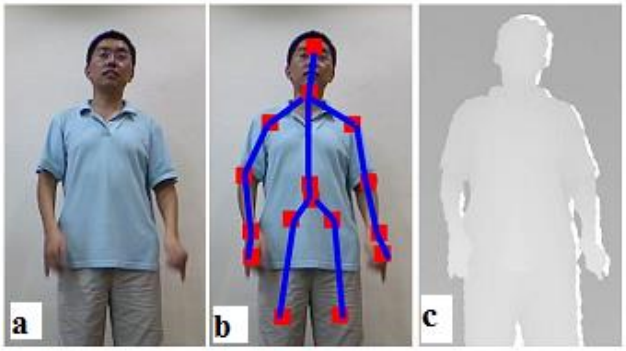


Figura 5- Exemplo de imagens adquiridas por Chao e colaboradores (2013). (a) imagens RGB, (b) esqueleto sobreposto na imagem RGB, (c) mapa de profundidade. (CHAO *et al.*, 2013).

Porfirio *et al.*(2013) fizeram o reconhecimento das 61 configurações de mãos da LIBRAS. Na aquisição das imagens, cinco pessoas pousaram para a câmera realizando uma determinada sequência de movimentos (Figura 6). Selecionando manualmente a visão frontal e lateral de cada CM, os autores construíram malhas 3D. Um total de 610 malhas 3D foram construídas. A segmentação da mão foi feita manualmente no *software* GNU *Image Manipulation Program* (GIMP). Na extração de características de cada malha aplicaram-se descritores esféricos harmônicos, que são insensíveis as operações de translação, rotação e escala. Para classificação são utilizadas máquinas de vetores de suporte com *Kernel* função de base radial (*Radial Base Function* - RBF) e *Kernel* Linear. Foram alcançadas taxas de reconhecimento de 85,68% para kernel RBF e 86,06% para Kernel Linear. Segundo os autores, o reconhecimento por malhas 3D mostrou-se promissor. Ressaltaram, porém, a necessidade de trabalhar com *hardware* auxiliar para a construção das malhas 3D de forma automática.

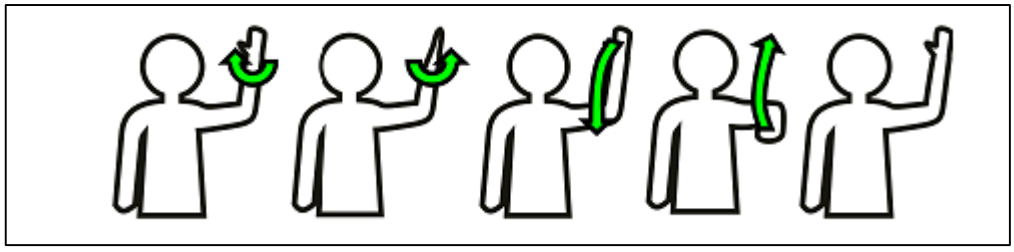


Figura 6- Sequência de movimentos adotada na aquisição das imagens. (PORFIRIO *et al.*, 2013).

Zhou *et al.*(2013) apresentaram um sistema de reconhecimento de 10 gestos (referentes aos caracteres numéricos de 0 a 9) usando sensor de profundidade Kinect®. O conjunto de dados foi coletado através da colaboração de 10 indivíduos. Cada indivíduo realizou 10 poses diferentes para o mesmo gesto. O estudo propõe como inovação uma medida de dissimilaridade usada para classificação dos gestos a *Finger-Earth Mover's Distance* (FEMD). A segmentação da mão foi facilitada pelo uso de uma pulseira preta no pulso e foi implementada por técnica de limiar. Após a etapa de segmentação, obtém-se a curva de série temporal, que contém propriedades topológicas da mão (ver Figura 7). A etapa de classificação é realizada através do casamento de modelos (*template matching*), onde cada gesto é rotulado em função da menor distância FEMD para um modelo. A taxa de classificação obtida foi de 93,9%.

Em (YUE e RUOYU, 2013), os autores utilizam a mesma base de dados construída por Zhou *et al.* (2013) . A segmentação foi baseada na biblioteca SDK do Kinect®. Após a detecção da mão, aplica-se uma transformação de coordenadas cartesianas para coordenadas polares nas imagens de profundidade. O autor não deixa claro qual classificador é empregado, mas sugere que SVM possui baixo custo computacional. Essa abordagem se mostrou superior quando comparado com trabalho de Zhou, pois a taxa de acerto foi de 97.1%.

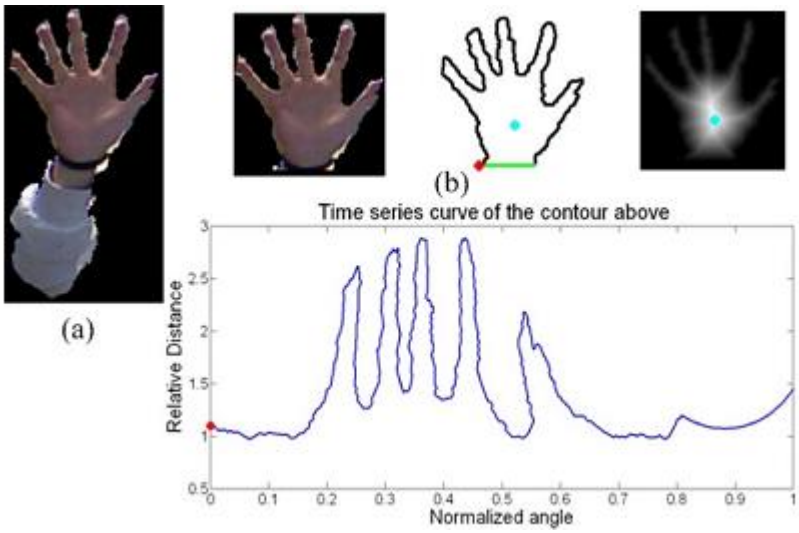


Figura 7- Extração de característica da postura da mão. (ZHOU *et al.*, 2013).

Através da análise dos trabalhos na área de reconhecimento de gesto das Línguas de sinais anteriormente citados, constatou-se que os estudos baseados em câmeras convencionais são sensíveis a problema de iluminação. A abordagem baseada em luvas e sensores permite obter uma boa taxa de reconhecimento. Entretanto, é necessário todo um aparato, que se mostra incômodo para o usuário, limitando os movimentos do mesmo. Assim, os sensores de profundidade despontaram como uma abordagem promissora, pelo fato de facilitar a segmentação da região de interesse, sendo mais apropriado para aplicações reais.

O Quadro 2 apresenta um resumo geral dos onze artigos revisados sobre reconhecimento de gestos. Nas colunas dessa tabela, para cada trabalho, apresenta-se os dispositivos de captura utilizados, o conjunto de imagens utilizado, o tipo de segmentação, as características empregadas, o algoritmo de classificação e os principais resultados.

Quadro 2- Sumário da Revisão Bibliográfica.

(Continua)

Autor	Dispositivo de captura	Materiais	Tipo de Segmentação	Extração de características	Algoritmo de Classificação	Taxa de reconhecimento
Ribeiro (2006)	Câmera Digital	10 posturas não associadas a Língua de Sinal	GMM e Limiar de cor de pele	Momentos de Hu e características geométricas	Casamento de modelos utilizando com medida de similaridade a distância euclidiana	94,43%
Maung (2009)	Câmera Digital	33 posturas do alfabeto da Língua de sinal de *Myanmar*	-	Orientação de Histograma Local	Redes Neurais	90%
Carneiro (2010)	Webcam	Colaboração de 6 usuários de LIBRAS formando conjunto de 7800 imagens RGB	Limiarização Simples	50 Descritores de Fourier/ Momentos Invariantes de HU	Redes Neurais	82,67%
Bragatto, Ruas e Castro (2006)	USB webcam	20 gestos estáticos do Alfabeto da LIBRAS	Redes neurais (MLP 3-3-6) filtros Morfológicos	Cada região de cor possui quatro características: coordenada do centroide, direção do eixo principal, razão de aspecto e coordenada do centro da palma da mão.	Redes neurais (MLP)	99,2%

(Continuação)

Autor	**Dispositivo de captura**	**Materiais**	**Tipo de Segmentação**	**Extração de características**	**Algoritmo de Classificação**		**Taxa de reconhecimento**
Maraqa *et al.* (2012)	Câmera Digital (indivíduo usando luva colorida)	1800 imagens do alfabeto Árabe (30 gestos)	Modelo de cor HIS	30 características (posições dos dedos relativas e as orientações em relação ao punho).	Redes Neurais de propagação direta	*Back-propagation*	79,33%
					Redes Neurais recorrentes	*Arquitetura Elman*	89,66%
						Arquitetura Jordan	84,54%
						Redes recorrentes puras	95,11%
Malassiotis e Strintzis (2008)	CCTV-color	20 posturas números (0 a 9) e letras do alfabeto da Língua de Sinais da Alemã.	Algoritmo Hierárquico (imagem de profundidade) Segmentação por cor (imagem RGB).	Descritores elípticos de *Fourier*	k Vizinhos mais próximos		RGB 83% Imagem de profundidade 81%
Yue e Ruoyu (2013)	Kinect®	NTU Dataset Zhou *et al.* (2013) ASL	Limiar de profundidade	Descritor de forma 2D	SVM		NTU 97% ASL 96,2%
Zhou *et al*. (2013)	Kinect®	10 gestos numéricos (0 a 9) da ASL	Limiar de profundidade e pulseira no pulso	Distância de Séries Temporais.	Casamento de modelos		98%

(Conclusão)

Autor	Dispositivo de captura	Materiais	Tipo de Segmentação	Extração de características	Algoritmo de Classificação		Taxa de reconhecimento
Chao *et al.* (2013)	Kinect®	1971 frases, incluído conjunto de imagem coloridas, mapa de profundidade e informações esqueleto.	Não implementa técnica de segmentação. Utiliza recursos do SKD Kinect®.	HOG+ informações do Kinect® (dados de Esqueleto)	SVM Latente		HOG 82,3% HOG+dados Esqueleto 86,0%
Rakun *et al.* (2013)	Kinect®	10 sinais SIBI com 12 amostras de cada sinal	Não implementa técnica de segmentação. Utiliza recursos da biblioteca SDK do Kinect®.	Propriedades geométricas (PG) de imagens de profundidade, DCTCC das imagens de cor e ângulos de quatro articulações do esqueleto (Skel)	GLVQ		Skel 87,92
							Propriedades Geométricas 79,58%
							DCTCC 85,29%
							Skel + PG 85,83%
							Skel + DCTCC 85,42
					Random Forest		Skel 85,92%
							PG 90,62%
							DCTCC 82.08%
							Skel + PG 94,37%
							Skel + DCTCC 82,08%
Porfirio *et al.* (2013)	Kinect®	5 pessoas participam da formação 610 malhas das configurações de mãos da LIBRAS	Segmentação manual	Descritores esféricos harmônicos	SVM	*Kernel* RBF	85,68%
						Kernel Linear	86,06%

3 REFERENCIAL TEÓRICO

Com o propósito de construir um sistema automatizado de reconhecimento das configurações de mãos de LIBRAS várias etapas ou métodos da visão computacional são usadas. Na Figura 8 mostra-se um diagrama em blocos do conjunto de etapas utilizadas nesse trabalho para o reconhecimento de configurações de mãos de LIBRAS. A seguir, descreve-se o embasamento teórico por traz de cada uma dessas etapas.

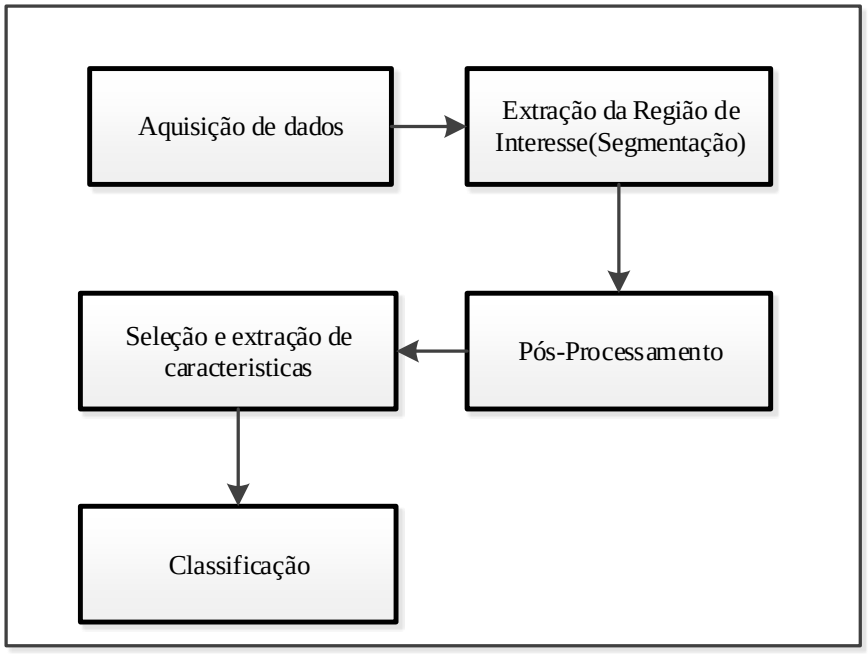


Figura 8- Diagrama de blocos do sistema de visão computacional seguido.

3.1 AQUISIÇÃO DE DADOS

Fisicamente, a etapa de aquisição de dados ocorre em um dispositivo que contém a interface de aquisição das imagens. Toma-se algumas decisões importantes que interferem no desempenho do sistema, como exemplo, posicionamento da câmera relativo ao usuário do sistema e gesto admitidos. Neste trabalho, o Kinect® da Microsoft® é o dispositivo responsável

pela aquisição de dados do sistema. É composto por três tipos de sensores: uma câmera RGB, uma câmera i*nfra-red* (IR) e sensores acústicos. Esse dispositivo gera tanto imagens RGB quanto imagens de profundidade. As imagens de profundidade, que mapeiam as distâncias de um objeto à câmera IR do dispositivo, são geradas através da utilização de um projetor de infravermelho associado à câmera IR. A câmera IR tem um limite prático para captura da radiação IR refletida. Apenas radiações de IR refletidas por objetos que se encontram a uma distância entre 0,8m e 3,5m são detectadas. A câmera gera sinais de vídeo a uma taxa de 30 quadros/segundo, com uma resolução de 640 × 480 *pixels*. O campo angular de visão é de 57 ° na horizontal e 43 ° na vertical (JUNGONG *et al*., 2013). Na Figura 9 mostra-se uma foto do sensor Kinect® utilizado.

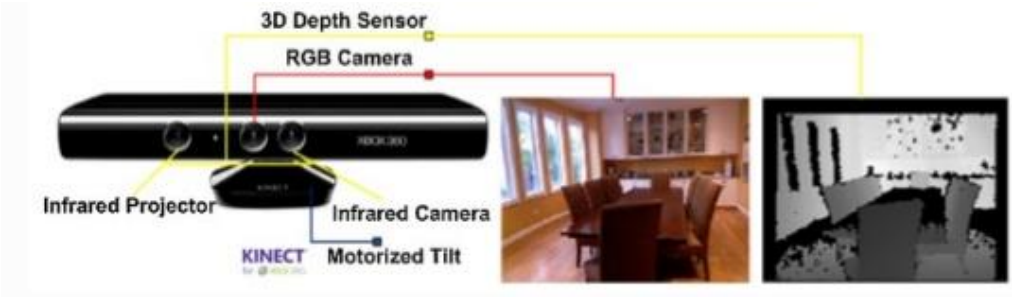


Figura 9- Sensor Kinect® e duas amostras de imagens capturadas simultaneamente, pela câmara RGB e pela câmara de profundidade. (JUNGONG *et al*., 2013).

3.2 EXTRAÇÃO DA REGIÃO DE INTERESSE

A extração ou segmentação da região de interesse (*Region of Interest – ROI*) é um importante passo do reconhecimento de gestos baseado em visão. Gonzalez e Woods (2002), afirmam que se trata de uma tarefa complexa na área de processamento de imagens, com forte interferência no sucesso da fase de classificação. Assim sendo, são necessários métodos robustos de segmentação da ROI.

Nesse trabalho, duas regiões de interesse são extraídas: a primeira delas corresponde a extração do braço e da mão do resto da imagem, enquanto que a segunda, a extração da mão (que é gesto propriamente dito). Essa seção trata de técnicas referentes a extração do braço e da mão do resto da imagem. A extração da mão (gesto) será apresentada na seção seguinte.

Diversos autores usam a técnica de limiar de profundidade como forma de segmentar o braço e a mão do resto da imagem (HASSANI *et al*., 2011; TANG, 2011; ZAFRULLA *et al*., 2011). A técnica é justificada pelo fato do tronco e do braço do indivíduo e do o fundo da imagem estarem, geralmente, em profundidades distintas em relação ao sensor de profundidade. No entanto, de acordo com Malassiotis e Strintzis (2008), o fato de que, em muitos casos, a mão do indivíduo não está suficientemente à frente do corpo, implica em dificuldade adicional para a aplicação da técnica de limiar.

Outra técnica possível de ser aplicada na segmentação baseia-se na utilização de técnicas de agrupamento, conforme proposto nesse trabalho. A seguir descreve-se essa técnica, utilizando o algoritmo de agrupamento *K-means*.

3.2.1 Algoritmo *K-means*

O algoritmo consiste em agrupar elementos em K grupos, em que K é quantidade de grupos que se deve informar ao algoritmo.

Seja $X= \{x_i\}$, $i=1,..., n$ o conjunto de n pontos a ser agrupado em K *clusters*: $C=\{c_k$ $k=1,.,K\}$. O algoritmo *K-means* encontra uma partição de tal forma que o erro quadrático entre o centro do *cluster* e os pontos desse *cluster* seja minimizado. Fazendo u_k ser a medida do centro do *cluster* c_k, o erro quadrático entre o centro do *cluster* u_k e os pontos pertencentes ao *cluster* c_k é definido como:

$$J(c_k) = \sum_{x_i \in c_k} \left| x_i - u_k \right|^2, \tag{1}$$

O objetivo do algoritmo é minimizar a soma do erro quadrático de todos os K *clusters* dado por:

$$J(C) = \sum_{k=1}^{K} \sum_{x_i \in c_k} |x_i - u_k|^2 \quad (2)$$

Na Figura 10 mostra-se os passos do algoritmo *K-means*. Na entrada são requeridos três parâmetros do usuário, que são o número de *cluster*s, K, a inicialização dos centros e a métrica de distância. A escolha mais crítica é a de K. Em geral o algoritmo é executado para diferentes valores de K e seleciona-se o valor que melhor agrupa os dados de forma correta (JAIN, 2010). Outra característica é a seleção dos centros inicias que levam a diferentes resultados em termos dos agrupamentos gerados, pois o algoritmo converge para mínimos locais. Com o objetivo de atingir um mínimo global para o erro médio quadrático, normalmente fixa-se o valor de K e realiza-se várias simulações, utilizando diferentes valores de inicialização dos centros. A melhor simulação é aquela que apresentar o menor erro quadrático. A métrica normalmente usada para algoritmo *K-means* é a distância Euclidiana, que calcula a distância entre os pontos e os centros dos *clusters*.

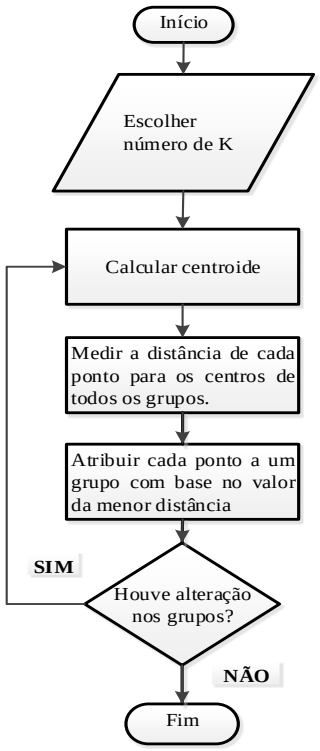


Figura 10- Passos do algoritmo *K-means.*

A Figura 11 mostra um exemplo da aplicação do algoritmo *K-means* para construção de 3 agrupamento de pontos no espaço bidimensional.

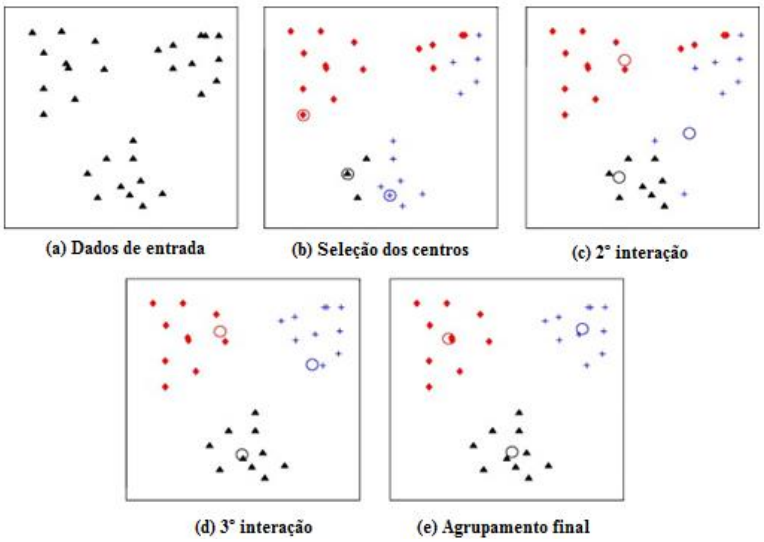


Figura 11- Ilustração K-*means*. a) pontos em duas dimensões; b) seleciona K=3; c) e d) interações; e) grupos resultantes formados pelo algoritmo. (JAIN, 2010).

Como o *K- means* é um algoritmo de baixa complexidade de programação, é possível que o mesmo seja aplicado em diversas situações, principalmente, quando o conjunto de dados é grande (THEODORIDIS e KOUTROUMBAS, 2008).

3.3 PÓS-PROCESSAMENTO

Nesse trabalho o pós-processamento é constituído de três etapas. Numa primeira etapa aplicam-se as seguintes transformações na imagem: filtragem de ruídos e rotação. Numa segunda etapa faz-se a extração da mão em relação ao antebraço utilizando a Transformada de Distância. Numa terceira etapa faz-se a padronização do tamanho das imagens e a normalização das profundidades.

Nessa seção descreve-se o referencial teórico relativo à determinação do ângulo de rotação da imagem, à rotação propriamente dita e à transformada de distância.

3.3.1 Momentos das Imagens

Os momentos de uma imagem permitem calcular determinadas propriedades geométricas de objetos presentes na imagem, como posição, tamanho e orientação. Para o cálculo desses parâmetros, utiliza-se uma imagem binária ou em escala de cinza bidimensional. Em geral, os momentos são quantidades numéricas que representam a soma de produtos de distâncias em relação a um referencial. (PROKOP e REEVES, 1992).

O momento m_{pq} de ordem $p + q$ da função de densidade é definido na equação (3):

$$m_{pq} = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} x^p y^q f(x, y) dx dy \qquad (3)$$

Para uma imagem binária MxN, a equação (3) transforma-se na equação (4):

$$m_{pq} = \sum_x \sum_y x^p y^q f(x,y) \tag{4}$$

Segundo Hu (1962), para representar todas as informações contidas num segmento de imagem é necessário o cálculo de um grande número de momentos. No entanto, para obtenção de algumas características geométricas, como área, centro de massa, orientação do eixo maior e menor da elipse circunscrita no objeto, apenas alguns momentos de baixa ordem são necessários.

3.3.1.1 Momentos de ordem zero: Área

Definido no espaço discreto como:

$$m_{00} = \sum_x \sum_y f(x,y), \tag{5}$$

m_{00} representa a área total do objeto segmentado da imagem.

3.3.1.2 Momentos de primeira ordem: Centro de Massa

Da equação (4), os momentos de primeira ordem, m_{10} e m_{01}, são usados para localizar as coordenadas (x', y') do centro de massa do objeto. Seus valores são dados por:

$$x' = \frac{m_{10}}{m_{00}} \tag{6}$$

$$y' = \frac{m_{01}}{m_{00}} \tag{7}$$

3.3.1.3 Momentos de segunda ordem

Os momentos de segunda ordem u_{02}, u_{20} e u_{11}, calculados a partir da equação (8):

$$u_{pq} = \sum_x \sum_y (x - x')^p (y - y')^q f(x, y), \tag{8}$$

são usados para determinar várias características dos objetos como a orientação e tamanho do eixo pricipal de um objeto. Os eixos principais são denominados de eixo menor e eixo maior. Em termos de valores, a orientação do eixo maior é dada pela equação (9) (PROKOP e REEVES, 1992).

$$\theta = \frac{1}{2} \tan^{-1}\left(\frac{2u_{11}}{u_{20} - u_{02}}\right), \tag{9}$$

em que θ corresponde ao menor ângulo entre o eixo maior e a horizontal (ver Figura 12) x' e y' são as coordenadas do centroide encontrados a partir das equações (6) e (7) respectivamente.

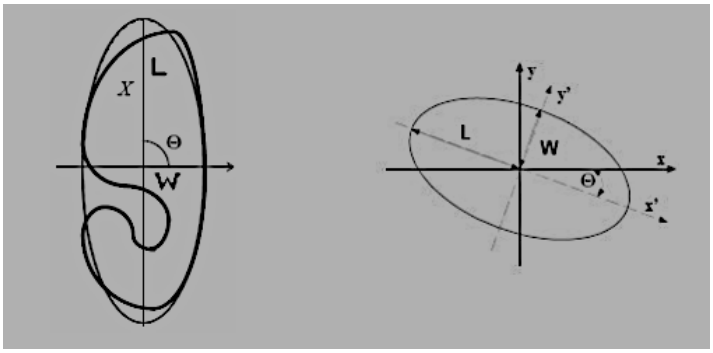


Figura 12- Elipse circunscrita no Objeto. À esquerda elipse que descreve a orientação de um contorno; à direita imagem mostrando eixo maior L o eixo menor W e a orientação θ. (RIBEIRO, 2006).

O comprimento do eixo maior W e eixo menor L da elipse são dados pelas equações (10) e (11), respectivamente. (DA FONTOURA COSTA e CESAR JR, 2010).

$$W = \left(\frac{2\left[u_{20} + u_{02} + \sqrt{(u_{20} - u_{02})^2 + 4u_{11}^2}\right]}{u_{00}} \right)^{1/2} \tag{10}$$

$$L = \left(\frac{2\left[u_{20} + u_{02} - \sqrt{(u_{20} - u_{02})^2 + 4u_{11}^2}\right]}{u_{00}} \right)^{1/2} \tag{11}$$

3.3.2 Rotação da Imagem

A rotação de imagens é um tipo de transformação geométrica que modifica a relação espacial entre os *pixels*. Em termos de processamento de imagens digitais a transformação geométrica consiste de duas operações básicas: (1) transformação espacial de coordenadas e (2) interpolação de intensidade que atribui valores de intensidade para *pixels* transformados (GONZALEZ e WOODS, 2002).

A primeira operação transforma uma imagem I para uma nova imagem I' modificando as coordenadas dos *pixels*. Quando alguma operação em cima das coordenadas de um *pixel* de uma imagem é realizada, o valor nem sempre é inteiro.

$$I(x, y) \rightarrow I^{'}(x^{'}, y^{'}) \tag{12}$$

A função de mapeamento do $R^2 \rightarrow R^2$ quase sempre tem um problema de re-amostragem. Isso ocorre quando a função de mapeamento tem como resultado coordenadas não inteiras.

Devido ao fato da primeira operação gerar coordenadas não inteiras, a segunda operação tem o objetivo de calcular um valor para o *pixel* da imagem resultante através de um processo de interpolação. São exemplos de interpolação comumente aplicadas nas imagens: interpolação pelo vizinho mais próximo (*Nearest Neighbor*), interpolação bilinear, interpolação bicúbica. A seguir é abordado a interpolação pelo vizinho mais próximo.

3.3.2.1 Rotação usando interpolação pelo vizinho mais próximo

A primeira operação da rotação é efetuada através das equações dadas em (13). Nessa operação, as coordenadas *x* e *y* são da imagem original, enquanto que as coordenadas *x´* e *y´* são da imagem final. Como os valores de *x´* e *y´* nem sempre são inteiros, utiliza-se o processo

de interpolação pelo vizinho mais próximo, com o objetivo de se escolher o pixel mais próximo na imagem de origem, para o qual o pixel *x*, *y* será mapeado.

$$\begin{aligned} x' &= x\cos(\theta) - ysen(\theta) \\ y' &= xsen(\theta) + y\cos(\theta) \end{aligned} \tag{13}$$

Em que:

θ = ângulo de rotação;

(x, y)= coordenadas dos *pixels* da imagem original;

(x', y')=coordenadas dos *pixels* da imagem rotacionada.

Na Figura 13 mostra-se um exemplo de rotação pelo vizinho mais próximo. Na Figura 13 (b) mostra que a imagem poderá ficar, nos limites de áreas continuas, com aspecto "serrilhado".

Figura 13- Ilustra uma operação de rotação. (a) imagem original; (b) imagem resultante de uma rotação de 21^0 no sentido horário usando interpolação pelo vizinho mais próximo (GONZALEZ e WOODS, 2002).

3.3.3 Transformada de Distância

Peixoto e Velho (2000), definem a transformada de distância T aplicada a um objeto O, como um campo escalar (ou vetorial) que representa distâncias mínimas entre o objeto e os pontos do espaço no qual ele está envolvido. A transformada T pode ser definida da seguinte maneira:

$$T(O) = \min_{pi \in o} dist(p, p_i), \quad (14)$$

em que p representa pontos do espaço e $dist$ representa uma função distância ou métrica utilizada. Assim, para cada ponto p do espaço, a transformada calcula a distância de p ao ponto p_i (pi pertence á borda O) que está mais próximo de p. A Figura 14 mostra a distância mínima entre um ponto p (interno ao objeto O) e a borda do objeto e a distância mínima entre um ponto q (externo) e o objeto.

A transformada de distância é uma operação que resulta como saída uma nova imagem cujos *pixels* representam a distância para uma determinada referência (BORGEFORS, 1986). Na Figura 15 (a) mostra-se um objeto em forma de F, enquanto que na Figura 15 (b) mostra-se a transformada de distância em relação a esse objeto F.

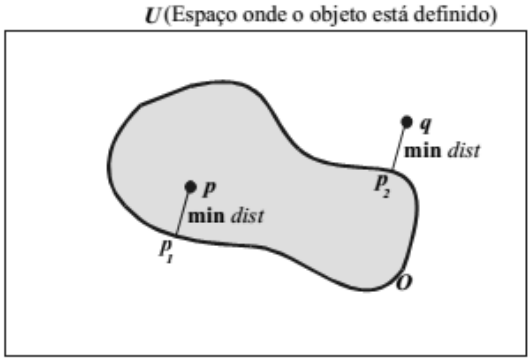


Figura 14- Mecanismo da Transformada de distância. (PEIXOTO e VELHO, 2000).

A imagem de saída do processo de transformação depende da métrica de distância usada. Como exemplo, na Figura 15(b) foi usada a distância Euclidiana.

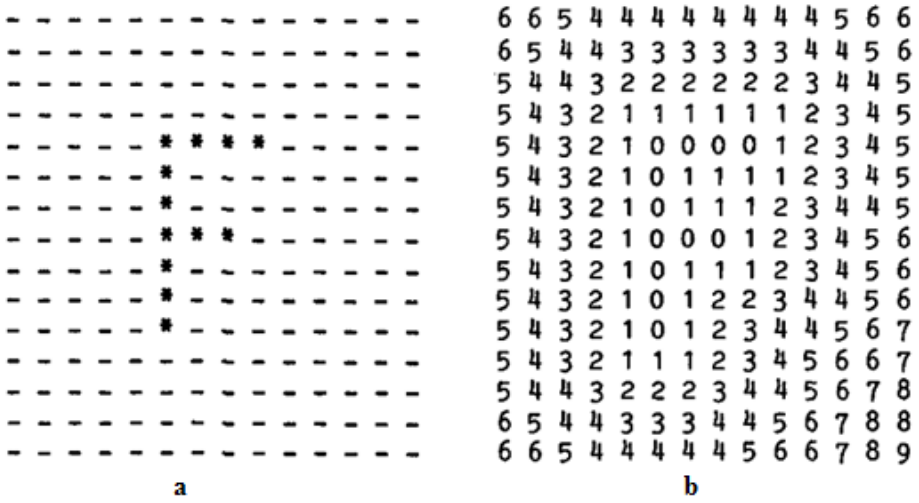


Figura 15- Exemplo de Transformada de distância. À esquerda é uma imagem binaria de um objeto em forma de F, enquanto que à direita a imagens resultante da transformada de distância. (BORGEFORS, 1986).

3.4 SELEÇÃO E EXTRAÇÃO DE CARACTERISTICAS

A seleção e a extração de características são etapas fundamentais que definem as variáveis de entrada do processo de classificação, levando em conta, principalmente, a capacidade dessas características melhor diferenciarem uma classe da outra. Nesse contexto, as técnicas de extração de características e redução de dimensionalidade como Análise de Componentes Principais (PCA), Análise Discriminante de Fisher (LDA); PCA bidimensional com redução em uma dimensão (2DPCA), PCA bidirecional com redução nas duas dimensões (2D2PCA), LDA bidirecional com redução em uma dimensão (2DLDA) e LDA bidirecional com redução nas duas dimensões (2D2LDA), assumem um relevante papel. Nesse trabalho utiliza-se para extração de características a técnica 2D2LDA. O embasamento teórico do método LDA e suas extensões como 2DLDA e 2D2LDA serão discutidos nessa seção.

3.4.1 Vetor de características

Uma imagem de um gesto pode ser representada por uma matriz A de dimensão mxn, em que m é o número de linhas e n o número de colunas de X. A mesma matriz também pode ser representada em forma de um vetor $\boldsymbol{a} = [a_{11}, a_{12,\ldots,} a_{mn}]^{\boldsymbol{T}}$ de dimensão $mnx1$, onde $a_{i,}$ é valor correspondente ao i-ésimo pixel. Na Figura 16 ilustra-se essa transformação de uma matriz em um vetor. Se considerar um conjunto de imagens, cada imagem pode ser representada com um vetor no espaço R^{mn}. Então, cada imagem é um ponto nesse espaço.

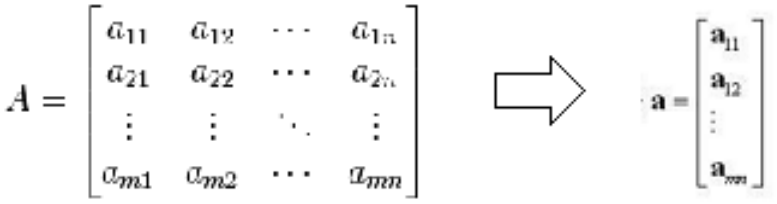


Figura 16- Representação de uma matriz-imagem no formato de vetor.

No reconhecimento de padrões, vetores de características que possuem valores redundantes não auxiliam na classificação das imagens (THEODORIDIS e KOUTROUMBAS, 2008). Portanto, é essencial, para um bom desempenho do classificador, a aplicação de técnicas que removem a redundância dos dados, como PCA, LDA e as técnicas derivadas das mesmas.

3.4.2 Análise de Discriminante de Fisher (FDA) e extensões

Análise de Discriminante de Fisher, também conhecida como Análise de Discriminante Linear (LDA), é uma técnica empregada para redução de dimensionalidade de dados pertencente a várias classes. O seu princípio é procurar uma direção ou um conjunto de direções

no espaço que possibilitem otimizar um critério de separabilidade das classes, ou seja, a projeção com LDA aumenta a distância entre as entre classes e reduz a distância interclasse.

Para uma amostra de treinamento na forma de vetores, a matriz de dispersão interclasse $S_{w,}$ é dada por:

$$S_w = \sum_{i=1}^{c} \sum_{j=1}^{N_i} (a_j^i - \bar{a}_i)(a_j^i - \bar{a}_i)^T \ , \tag{15}$$

em que: a_j^i é o i-ésimo vetor de treinamento (elemento) da classe i, $\bar{a}_\iota$ é a média dos elementos da classe i, c é o número total de classes e N_i é o número de elementos da classe i. A matriz de dispersão entre as classes S_b, é dada por:

$$S_b = \sum_{i=1}^{c} N_i (\bar{a}_i - \bar{a})(\bar{a}_i - \bar{a})^T , \tag{16}$$

em que: $\bar{a}$ é média global das amostras de treinamento.

Na técnica LDA, procura-se maximizar o critério de separação de classes dado pela equação (17). Pode-se mostrar que a matriz de projeção X que maximiza o esse critério é aquela formada pelos autovetores que correspondem aos maiores autovalores da matriz $S_w^{-1} S_b$. A dimensão da matriz X depende do número de dimensões que se deseja no vetor projetado. Um vetor projetado a_p é obtido a partir da equação (18).

$$j(X) = \frac{\left| X^T S_b X \right|}{\left| X^T S_w X \right|} \tag{17}$$

$$a_p = X^T a, \tag{18}$$

Em que a matriz de projeção X= $[u_1, u_2, \dots, u_q]$ é composta dos q autovetores associados aos q maiores autovalores $\lambda_1, \lambda_2 \dots \lambda_q$ não nulos da matriz $S_w^{-1} S_b$ e a é imagem original.

3.4.2.1 2DLDA

Na implementação do método de redução de dimensionalidade 2DLDA, reduz-se uma das dimensões de uma matriz bidimensional. No nosso estudo, essas matrizes correspondem a imagens. As resoluções das imagens podem variar muito, dependendo da qualidade da imagem que se deseja estudar. Numa imagem discreta de 640 × 480 *pixels*, o total de *pixels* da matriz que a representa é igual a 307.200 *pixels*. Se utilizássemos a técnica LDA de redução de dimensionalidade, a matriz de dispersão $S_w^{-1}S_b$ seria muito grande e o cálculo dos autovetores envolveria um esforço computacional elevado.

A técnica 2DLDA proposta por Li e Yuan (2005), para redução de dimensionalidade não exige a conversão de espaço R^2 para R^1, como a técnica LDA. Como resultado a matriz de dispersão é bem menor do que a do método LDA.

Sejam: c o número de classes, N o total de amostras de treinamento, Ni o número de amostras da classe i, $A_j^{(i)}$ a j-ésima imagem da classe i de dimensão mxn, $\bar{A}^{(i)}$ a média das imagens da classe i, e $\bar{A}$ a média total das imagens. Usando as imagens de treinamento, as matrizes de dispersão entre classes S_B e a matriz de dispersão interclasse, S_w são dadas por:

$$S_b = \frac{1}{N}\sum_{i=1}^{c} N_i (\bar{A}_i - \bar{A})^T (\bar{A}_i - \bar{A}) \tag{19}$$

e

$$S_w = \frac{1}{N}\sum_{i=1}^{c}\sum_{j=1}^{N_i} (A_j^{(i)} - \bar{A^{(i)}})^T (A_j^{(i)} - \bar{A^{(i)}}) . \tag{20}$$

No método 2DLDA procura-se obter uma matriz H que maximiza o critério de separação de classes dado pela equação (21) (critério de Fisher).

$$J(H) = \frac{\left|H^T S_b H\right|}{\left|H^T S_w H\right|}. \tag{21}$$

A matriz H= $[h_1, h_2, \dots, h_q]$ formada pelos autovetores da matriz $S_w^{-1}S_b$, associados aos seus q maiores autovalores é solução que maximiza o critério dado pela equação (21).

A matriz projetada nesse caso é dada por:

$$B = AH, \tag{22}$$

em que B possui dimensão mxq. A redução de dimensionalidade é obtida considerando que $q < n$. Quanto menos autovetores h_i forem utilizados na montagem de H, maior a redução da dimensionalidade na direção horizontal.

3.4.2.2 2D2LDA ou 2DLDA Bidirecional

O método 2D2LDA proposto em (NOUSHATH, KUMAR e SHIVAKUMARA, 2006), tem por objetivo fazer reduções na direção horizontal e vertical de uma matriz ou imagem.

A técnica consiste em encontrar duas matrizes de projeções: X= $[x_1, x_2, \dots, x_d]$ e Z= $[z_1, z_2, \dots, z_q]$. A matriz X corresponde a matriz de projeção obtida pela técnica 2DLDA. Através da matriz X obtém-se a redução da dimensionalidade na direção horizontal. Para obtenção da matriz de projeção Z, são determinadas as matrizes de dispersão do método "*alternative*" 2DLDA, designado assim pelos autores, como forma de redução na direção vertical da imagem de entrada. As matrizes de dispersão S_b e S_w são:

$$S_b = \frac{1}{N}\sum_{i=1}^{c} N_i (\bar{A}_i - \bar{A})(\bar{A}_i - \bar{A})^T \tag{23}$$

$$S_w = \frac{1}{N}\sum_{i=1}^{c}\sum_{j=1}^{N_i} (A_j^{(i)} - \bar{A^{(i)}})(A_j^{(i)} - \bar{A^{(i)}})^T \ . \tag{24}$$

Da mesma forma, calcula-se os autovetores de $S_w^{-1}S_b$ que são empregados como projeção da imagem original. A matriz $S_w^{-1}S_b$ possui dimensão mxm. A matriz X é de nxd e Z

de dimensão mxq. A matriz de características Ck é dimensão dxq, é obtida projetando-se a imagem Ak nas matrizes X e Z simultaneamente (Ver equação (25)).

$$C_k = Z^T A_k X \tag{25}$$

3.5 CLASSIFICAÇÃO

Após a extração de características, a última etapa do sistema de reconhecimento de padrões é a classificação. A seguir descreve-se o método de classificação utilizado nesse trabalho, ao algoritmo *k- Vizinhos mais próximos*.

3.5.1 Algoritmo k-Vizinhos mais Próximos (k *nearest neighboor* – kNN).

O algoritmo denominado de regra do vizinho mais próximo (*Nearest Neighbor* – NN) tem como princípio de funcionamento encontrar o vizinho mais próximo de um dado vetor de características desconhecido, x. Já no algoritmo (*k-Nearest Neighbor* – kNN) são encontrados os k vizinhos mais próximos do padrão de teste. Portanto, esse método tem fácil implementação e é uma técnica muito utilizada no reconhecimento de padrões.

A fase de treinamento do algoritmo consiste em armazenar os padrões de treinamentos com suas classes correspondentes. Com isso, o kNN classifica uma dada instância baseado nas respectivas classes dos k vizinhos mais próximos da base de treinamento. Resumidamente, dado um padrão, o algoritmo calcula a distância para cada padrão de treinamento e ordena os elementos da base de treinamento do mais próximo ao mais distante. Após essa ordenação, escolhe-se os k primeiros e verifica-se a que classes os mesmos pertencem.

A regra de classificação, a função que calcula a distância entre dois pontos e a escolha do valor de k, são três parâmetros importantes no método kNN. A regra de classificação diz respeito à relevância de cada um dos k elementos selecionados. A função de distância mede a

distância no espaço multidimensional. A escolha do valor de *k* permite escolher qual a fronteira na vizinhança do padrão a ser classificado a ser utilizada na classificação.

Em (WANG,2006), são apresentadas duas regras de classificação clássicas: maioria na votação e peso pela distância. Na votação, cada elemento tem uma influência igual e a classe atribuída a um padrão que se deseja classificar é aquela mais frequente entre os k padrões utilizados na classificação. Na segunda regra, cada k vizinho tem um peso inversamente proporcional a sua distância. A Figura 17 ilustra o kNN para caso de k=11.

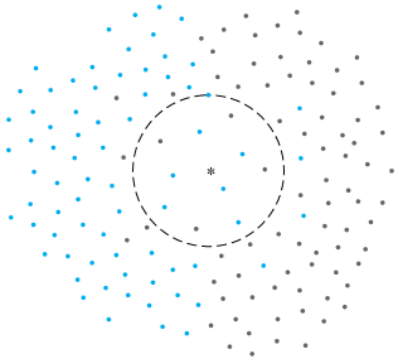


Figura 17- O ponto estrela sendo classificados na classe azul, pois dentre os 11 vizinhos mais próximo a classe azul é mais frequente nos padrões de treinamento com 7 pontos e classe preta com 4 pontos. (THEODORIDIS e KOUTROUMBAS, 2008).

Funções de distância são necessárias em muitos algoritmos modernos. As medidas de distância de uma maneira geral podem ser definidas como medidas de similaridade, e dissimilaridade. A primeira é para definir o grau de semelhança entre as instâncias e realizam o agrupamento de acordo com a sua coesão, e a segunda mede as diferenças dos atributos das instâncias. Uma variedade de funções de distância está disponível na literatura, incluindo a Euclidiana, *Hamming*, *Minkowsky*, *Mahalanobis*, *Camberra*, *Chebychev*, *Manhattan*. A seguir serão abordas duas métricas usadas nesse trabalho.

3.5.1.1 Distância Euclidiana

A distância Euclidiana entre os indivíduos i e j é dada analiticamente por:

$$D_{ij} = \sqrt{\sum_{k=1}^{n} \left(p_{ki} - p_{kj}\right)^2}\,, \tag{26}$$

em que: $n = 1,2,\cdots,k$;

p_{ki} = valor da variável k para o individuo i;

$\boldsymbol{p_{kj}}$ = valor da variável $\boldsymbol{k}$ para o individuo $\boldsymbol{j}$.

3.5.1.2 Distancia *Manhattan* ou *City Block*

A distância *Manhatan* entre os indivíduos i e j é dada analiticamente por:

$$D_{ij} = \sum_{k=1}^{n} \left|p_{ki} - p_{kj}\right|, \tag{27}$$

em que:

$\boldsymbol{n = 1, 2, \cdots, k}$;

$\boldsymbol{p_{ki}}$ = valor da variável $\boldsymbol{k}$ para o individuo $\boldsymbol{i}$;

$\boldsymbol{p_{kj}}$ = valor da variável $\boldsymbol{k}$ para o individuo $\boldsymbol{j}$.

ou seja, é distância entre dois pontos é a soma das diferenças absolutas de suas coordenadas.

Essas funções oferecem diferentes resultados na classificação, pois como podemos ver na Figura 18 a distância Euclidiana seria o segmento de uma reta e a distância *Manhattan* seria um segmento de retas na vertical quanto na horizontal.

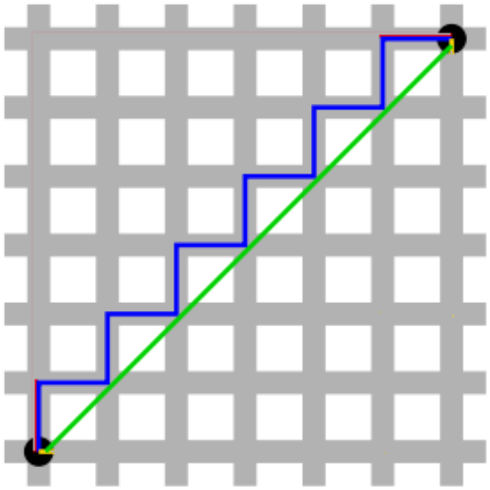


Figura 18- Comparação entre a distância de *Manhattan* (em azul) e a distância Euclidiana (em verde).

4 MATERIAS E MÉTODOS

Para o desenvolvimento do sistema de reconhecimento das 61 configurações de mãos (CM) de LIBRAS, em imagens de profundidade provenientes do sensor Kinect®, utilizou-se um computador com as seguintes configurações de *hardware*: processador Intel® Core (TM) i3-2310M 2.10GHz, 4GB de memória RAM e 500GB de HD. O ambiente de simulação foi o Matlab2013®. Na Figura 19 é ilustrado um diagrama em blocos do sistema de reconhecimento de padrões desenvolvido. A seguir descreve-se como foram realizadas cada uma das etapas mostradas nesse diagrama em blocos.

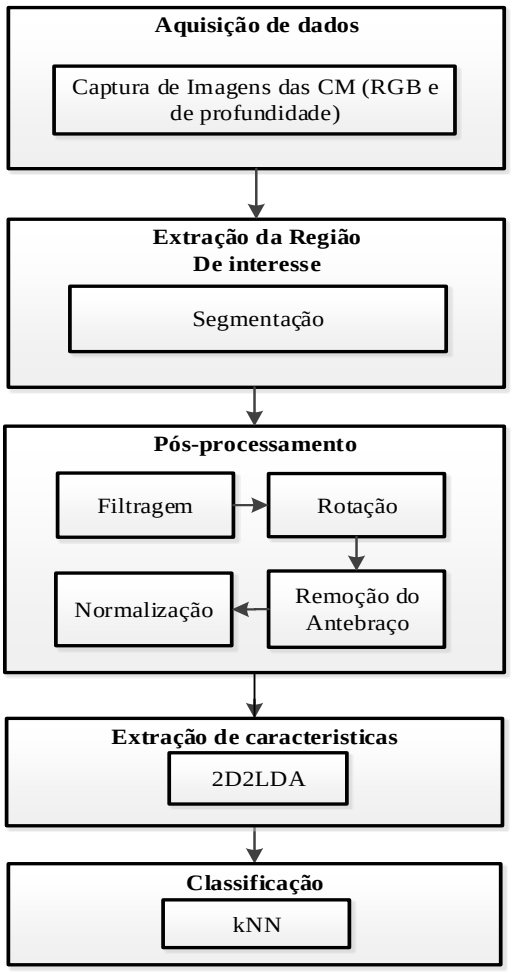


Figura 19- Sistema de reconhecimento de padrões.

4.1 AQUISIÇÃO DAS IMAGENS DE PROFUNDIDADE

Nesta etapa foram utilizados o sensor de profundidade Kinect® e o Microsoft® Kinect SDK versão 1.8 para obter imagens de profundidade com resolução de espacial de 640x480. Também foram adquiridas imagens RGB.

As imagens foram adquiridas nas dependências da Escola Estadual do Amazonas Augusto Carneiro Dos Santos, com a devida autorização formal da diretoria da Escola. Essa atividade de aquisição das imagens para a construção do conjunto de dados utilizado na presente dissertação foi realizada em conjunto com o mestrando Robson Silva de Souza que também desenvolveu dissertação dentro da mesma temática (SOUZA, 2015).

O ambiente de aquisição das imagens tinha iluminação artificial e fundo na frente do qual o indivíduo se posicionava para a captura da imagem era homogêneo. Apesar desta última condição, não ser necessária quando se trabalha com imagens de profundidade, a mesma visou facilitar a segmentação das configurações de mão no caso de uso das imagens RGB.

O banco de imagens foi formado por 12200 imagens resultantes da seguinte combinação: 10 indivíduos, 61 configurações de mão, 20 imagens por configuração, por indivíduo. Dentre os 10 indivíduos, duas são mulheres e oito são homens. Sete possuem LIBRAS como primeira língua (como resumido na Figura 20). As idades dos voluntários variam entre 18 e 25 anos. Na seleção dos voluntários procurou-se uma diversidade em termos do tamanhos das mãos: maõs pequenas, mãos médias e mãos grandes. Os usuários foram incentivados a rotacioanar e transladar suas mãos lentamente num intervalo entre 45^o e 135^o. A câmera foi, fisicamente, posicionada a 1,4 m do fundo e usuário ficou a uma distância média de 1,3 m (Figura 21).

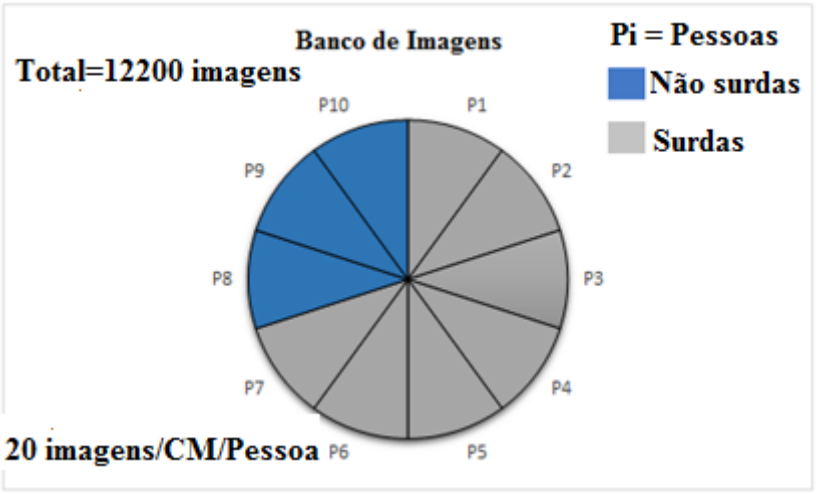


Figura 20-Ilustra a composição do banco de dados.

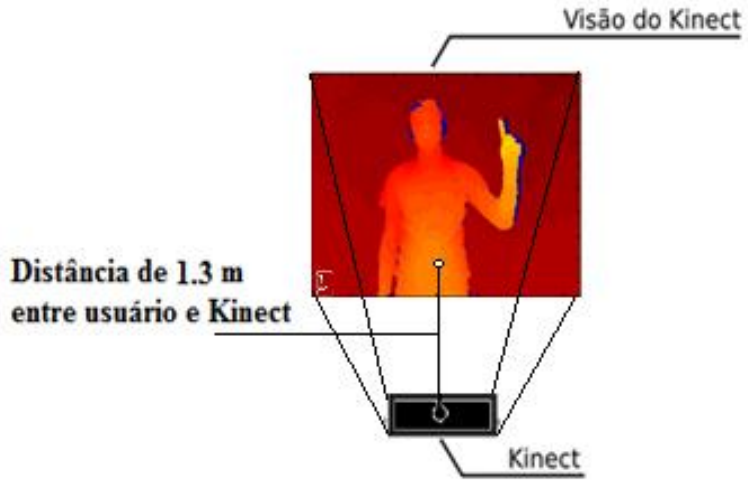


Figura 21- Posicionamento dos indivíduos durante o processo de aquisição das imagens.

Na Figura 22 a seguir são apresentados exemplos de imagens do banco nas duas modalidades: RGB e de profundidade.

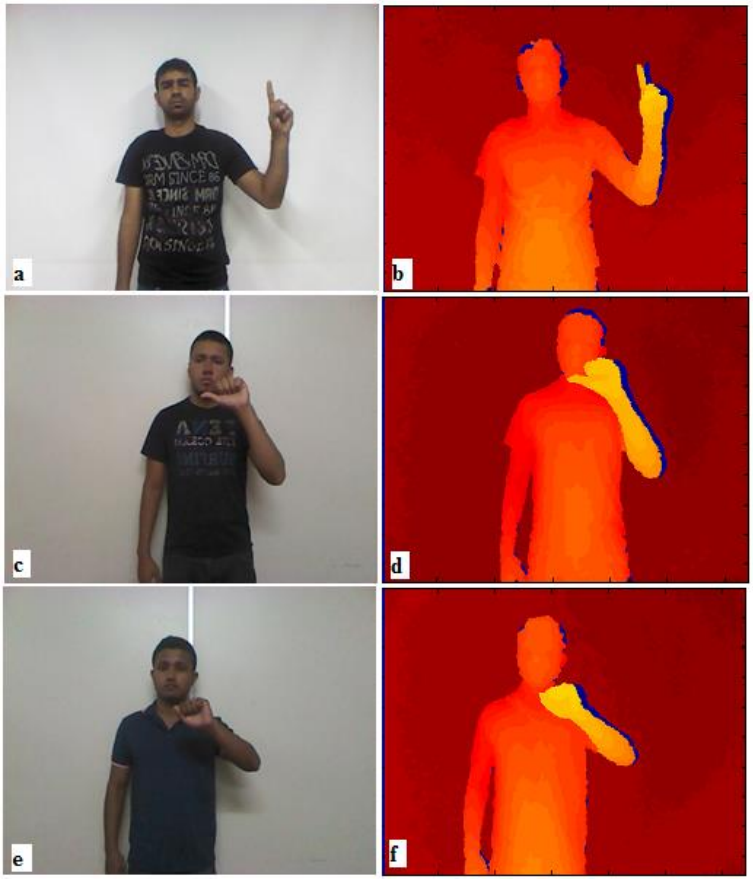


Figura 22- Exemplos de imagens que constitui o banco de imagens (a), (c) e (e) imagens RGB e (b), (d) e (f) imagens de profundidade respectivas.

4.2 EXTRAÇÃO DA REGIÃO DE INTERESSE

Uma vez que as imagens das configurações de mãos já foram adquiridas, a primeira etapa do trabalho consiste na segmentação da mão e do antebraço nas imagens de profundidade.

4.2.1 Segmentação

Neste trabalho utiliza-se o sensor de profundidade que elimina severos empecilhos na cena do ambiente como luminosidade, cor de pele do usuário e *background*. Com isso, conforme já relatado em trabalhos revisados anteriormente, ocorre uma facilitação da etapa de segmentação.

No trabalho atual foi usada a técnica de limiar de Otsu (1975) com objetivo de subtrair o fundo que, presumidamente, é separado do corpo do usuário (Figura 23). Na técnica que se utiliza para segmentação e que será descrita a seguir, a eliminação do fundo é extremamente importante, pois evita a formação de agrupamentos diferentes para cada tipo de fundo.

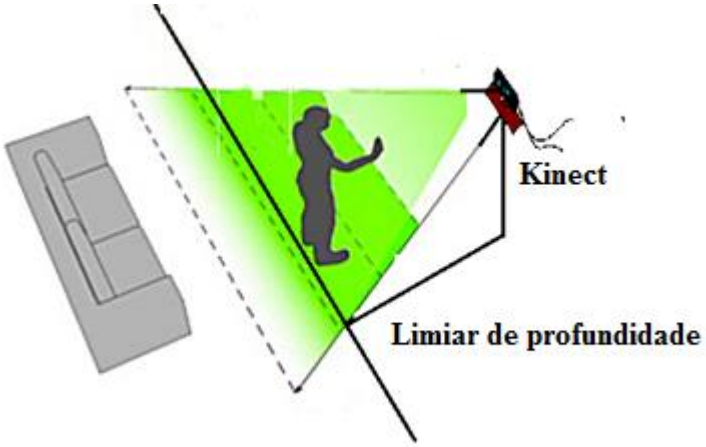


Figura 23-Ilustra metodologia adotada.

As Figura 24(a) e (b) mostram imagens plotadas com Matlab®. Nesse *plot* os *pixels* maiores que o limiar de profundidade são levados a zero (representados pela cor azul).

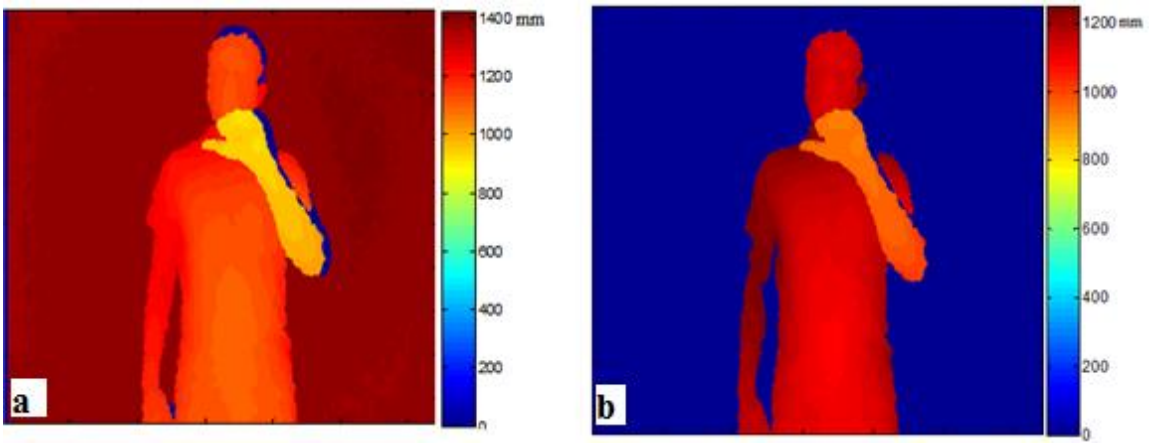


Figura 24- Remoção do fundo. Indivíduo P_1, Configuração CM_1: (a) mostra que os valores do fundo estão em 1.4 m (área em vermelho escuro), (b) mostra a imagem resultante da subtração do fundo.

Para a segmentação da região de interesse, propriamente dita, mão e antebraço, usou-se o algoritmo *K-means* visto na seção 3.2.1. A utilização desse algoritmo baseou-se na observação de que as diversas partes do corpo formam *grupos* compactos quando analisadas

em função da dimensão z (profundidade) da imagem. Os passos e os parâmetros escolhidos para aplicação desse método de segmentação são descritos a seguir:

1 Transforma-se a imagem I, em formato de matriz, para a forma de vetor x.

2 Inicializa-se K=3, 4 ou 5 para que sejam formados os grupos.

3 Inicializa-se os centros iniciais com $C_1 = 0$, C_2=máximo valor do mapa de profundidade, C_3=mínimo valor do mapa de profundidade, C_4= média entre o máximo e o mínimo e C_5=$C_4/2$.

4 Adota-se o quadrado da distância Euclidiana como métrica de distância para formação dos agrupamentos. Ver equação (28).

$$D_{xi\,c} = (x_i - C_i)^2, \tag{28}$$

O *cluster* referente à região de interesse é selecionado como aquele com o centro mais próximo da câmera.

4.3 PÓS-PROCESSAMENTO

O pós-processamento é constituído de três etapas. Numa primeira etapa aplicam-se as seguintes transformações na imagem: filtragem de ruídos e rotação. Numa segunda etapa faz-se a extração da mão em relação ao antebraço utilizando a transformada de distância. Numa terceira etapa faz-se a padronização do tamanho das imagens e a normalização das profundidades.

4.3.1 Filtragem

Na maioria das CM segmentadas, a mão e o antebraço situam-se à frente do tronco e a segmentação ocorre sem problemas. Quando o antebraço do usuário não está posicionado mais à frente do tronco, e sim lateralmente, a região do antebraço e partes do corpo, como a cabeça, podem ser segmentados no mesmo grupo. Diz-se, nesses casos, que a segmentação apresenta ruídos. Na Figura 25 mostra-se um exemplo em que o antebraço é segmentado juntamente com a cabeça.

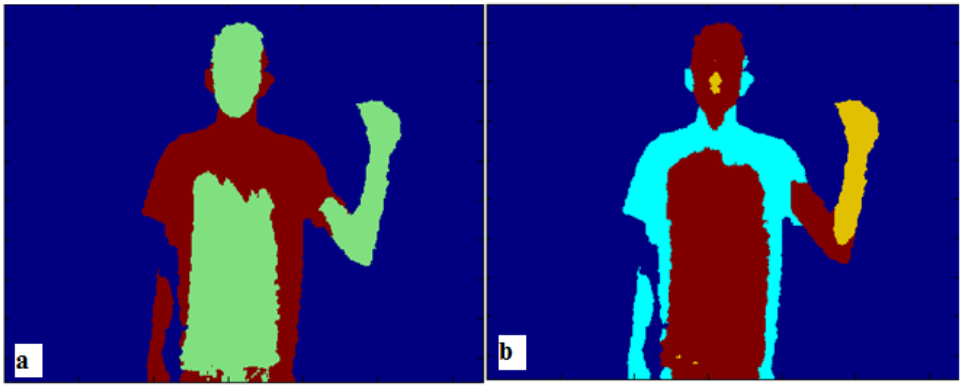


Figura 25- Configuração da mão com antebraço situado lateralmente ao corpo. (a) resultado do método *K-means* com 3 grupos, mostrando o antebraço segmentado em um mesmo grupo com a cabeça e a barriga. (b) resultado do método *K-means* com 4 grupos mostrando o antebraço segmentado em um mesmo grupo com a cabeça.

Para filtragem dos ruídos, aplica-se o seguinte procedimento: são medidas as distâncias dos centros dessas regiões segmentadas para o eixo vertical do corpo do usuário. Assumindo que as regiões não estão conectadas entre si, a região contendo o antebraço e a mão (Região de Interesse - ROI) é aquela que está mais distante do eixo central do usuário. A ROI selecionada na Figura 26(a) aplicando esse critério é mostrada na Figura 26(b).

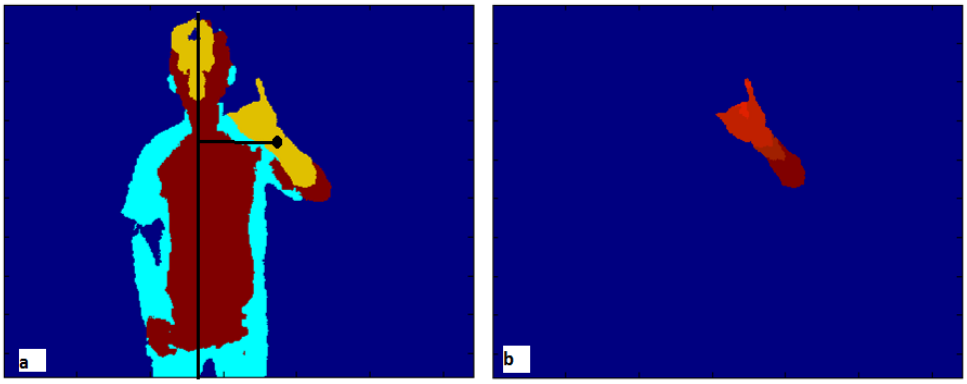


Figura 26- Filtragem. Indivíduo P_2, configuração CM_3. (a) ilustra que a região do braço está mais distante do eixo vertical do que a região da cabeça. (b) mostra o resultado da filtragem.

4.3.2 Rotação

A seção 3.3.1 apresentou um método para calcular a orientação de um objeto, enquanto que a seção 3.3.2 descreve o processo de rotação da imagem. Essas operações são efetuadas em sequência. A rotação é importante para padronizar todas as imagens numa mesma orientação, a direção vertical. Essa padronização facilita a aplicação do algoritmo de extração da mão e viabiliza a extração de características que serão vistos nas seções seguintes.

Todas as imagens são rotacionadas em relação ao eixo vertical, de tal forma que as mesmas fiquem alinhadas com o mesmo. Nas Figura 27(b) e (d) mostram-se exemplos de ROI rotacionadas referentes as Figura 27(a) e (c), respectivamente.

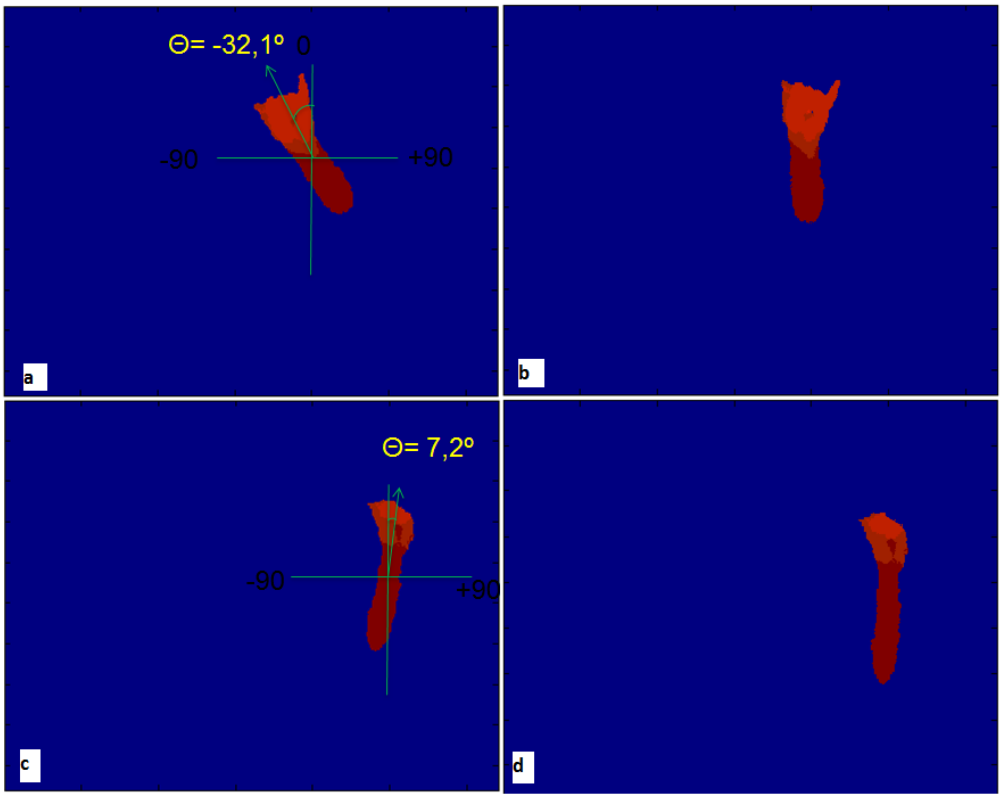


Figura 27- Rotação P_7, CM_3. (a) e (c) mostram a orientação da CM com relação ao eixo vertical, (b) e (d) resultado da rotação de (a) e (c), respectivamente.

4.3.3 Remoção do antebraço

Devido a não uniformidade na parte inferior da região de interesse, dependendo do resultado da segmentação, o seccionamento do antebraço pode ser efetuado em até duas etapas.

Na primeira etapa os seguintes passos são obedecidos: determina-se o centroide da imagem e traçam-se os eixos maior e menor (descritos na seção 3.3.1) passando pelo mesmo (Figura 28(a)); calcula-se a razão entre o eixo maior e o eixo menor; verifica-se se essa razão é maior do que 3,3; se for maior, elimina-se a parte do antebraço abaixo do centroide e aplica-se a transformada de distância; se for menor aplica-se simplesmente a transformada de distância. A transformada de distância será detalhada na segunda etapa. O objetivo da primeira etapa é verificar se o cotovelo foi segmentado junto com o antebraço. Caso tenha sido, o mesmo é eliminado. Na Figura 29 mostra-se um exemplo em que o cotovelo é segmentado junto com

o antebraço. Quando o cotovelo é segmentado junto com o antebraço, a razão entre o eixo maior e o eixo menor é maior do que 3,3. Tal valor foi obtido experimentalmente a partir da análise do conjunto de imagens do banco. O cotovelo é eliminado, removendo da ROI a parte do antebraço abaixo do centroide (Figura 28(b)).

A segunda etapa objetiva a extração da mão a partir da ROI. Para esse fim, utiliza-se a transformada de distância utilizada por Deimel e Schröter (1998)(seção 3.3.3) com o objetivo de determinar o centro e o raio da palma da mão. A transformada de distância consiste em associar a cada pixel p do objeto a menor distância euclidiana, $D(p,q)$, de p para um pixel de borda q (Figura 28(c)) encontrada a partir da equação (29). Gera-se então uma imagem I (Figura 28(d)) em que a intensidade de cada pixel é proporcional a essa distância. As coordenadas do centro, (x_{centro}, y_{centro}) corresponde às coordenadas do pixel de maior intensidade em I. O raio R da palma da mão corresponde ao valor de nível de cinza desse pixel, conforme mostrado na equação (30). A coordenada x_{corte} do ponto de corte (x_{corte}, y_{corte}) utilizado para extração da mão (Figura 28 (e)) é encontrada multiplicando-se o valor de R por um escalar igual a 1,53 (DEIMEL e SCHRÖTER, 1998) e somando-se com a coordenada x_{centro} da palma da mão, conforme mostrado na equação (31). A região que estiver abaixo do ponto x_{corte} é eliminada da ROI. Na Figura 28(f) mostra-se o resultado da extração da mão aplicando essa transformação.

$$B = I - (I \ominus S), \tag{29}$$

em que:

$S = \begin{bmatrix} 0 & 1 & 0 \\ 1 & 1 & 1 \\ 0 & 1 & 0 \end{bmatrix}$ é o elemento estruturante;

$I =$ imagens original;

B=imagem da borda.

$$R = \max(I) \tag{30}$$

$$x_{corte} = x_{centro} + 1{,}53R \tag{31}$$

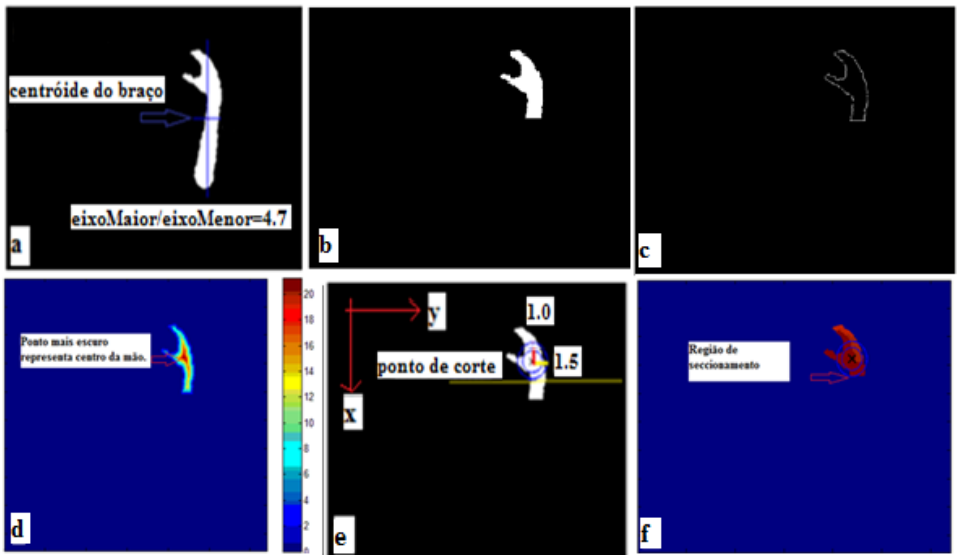


Figura 28- Descrição das fases da primeira etapa da extração da mão. Indivíduo P_1, configuração CM_{29}. (a) ilustra o eixo menor α e eixo maior β e o centro da região (interseção entre os eixos), (b) ilustra a remoção da parte inferior ao centro, (c) ilustra borda do objeto, d) é imagem de distância gerada a partir da TDE, (e) ilustra centro da palma da mão e ponto de corte da segunda fase, (f) mostra resultado da segunda etapa.

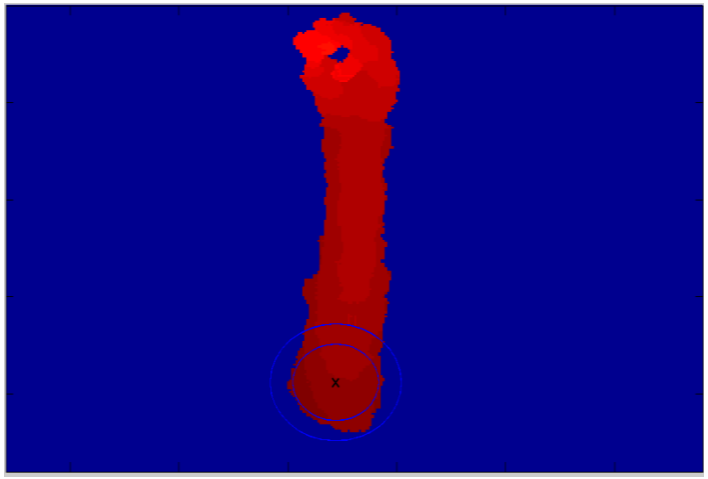


Figura 29- Exemplo de Falso centro. Indivíduo P_1, CM_{27}. Região do cotovelo interfere no método adotado para remoção do antebraço.

A Figura 30 ilustra cada uma das etapas da extração da mão para o indivíduo P_1, configuração CM_3.

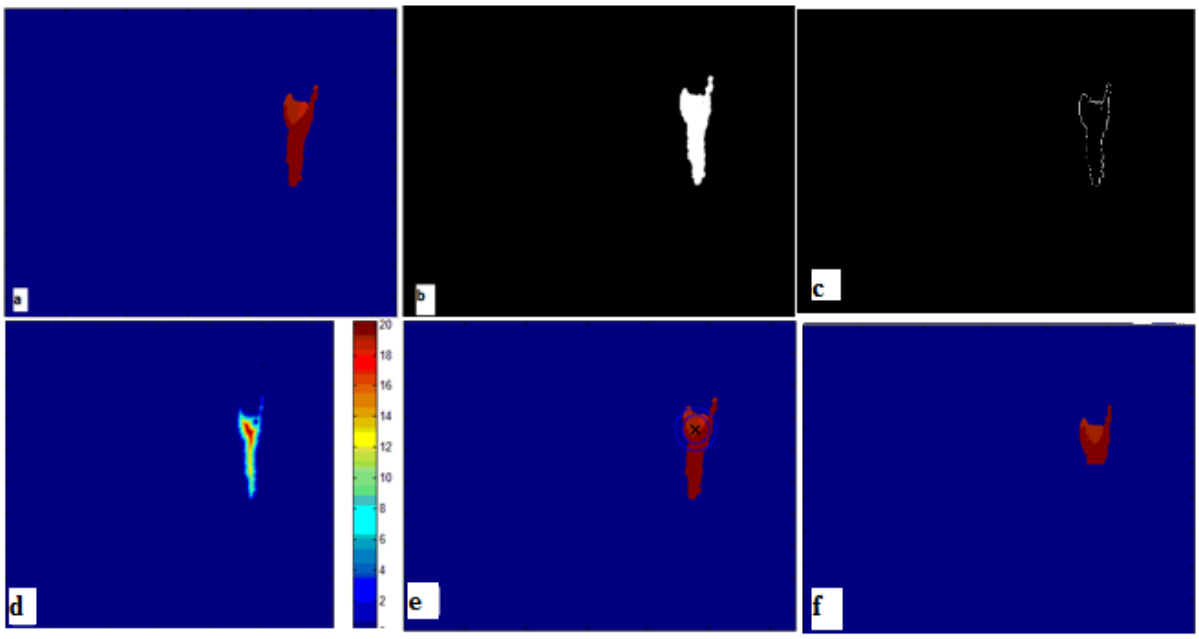


Figura 30- Processo de segmentação da mão. Indivíduo P_1, CM_3. (a) região do braço segmentada, (b) mascara binária, (c) borda da máscara binária, (d) imagem de distância gerada após aplicação da TDE, (e) ilustra circunferência (menor) que é uma representação aproximada da circunferência da palma da mão e circunferência (maior) que intercepta a região de corte, (f) resultado da segmentação.

4.3.4 Padronização

A padronização é ainda uma sub-etapa do pós-processamento, visando a adequação da imagem para a etapa de extração de características. O conjunto de imagens foi padronizado com dimensão 130x134, que corresponde as maiores dimensões encontradas nas imagens segmentadas. Imagens menores são preenchidas linhas e colunas de zero abaixo e à direita. Na Figura 31(a) mostra-se essa operação de *crop* (menor retângulo que o objeto de interesse ocupa) de uma imagem cuja dimensão obtida é de 85x51 *pixels* (Figura 31(a)) e que foi preenchida com zeros (representados pela cor azul), gerando a imagem com tamanho 130x134 mostrada na Figura 31(b).

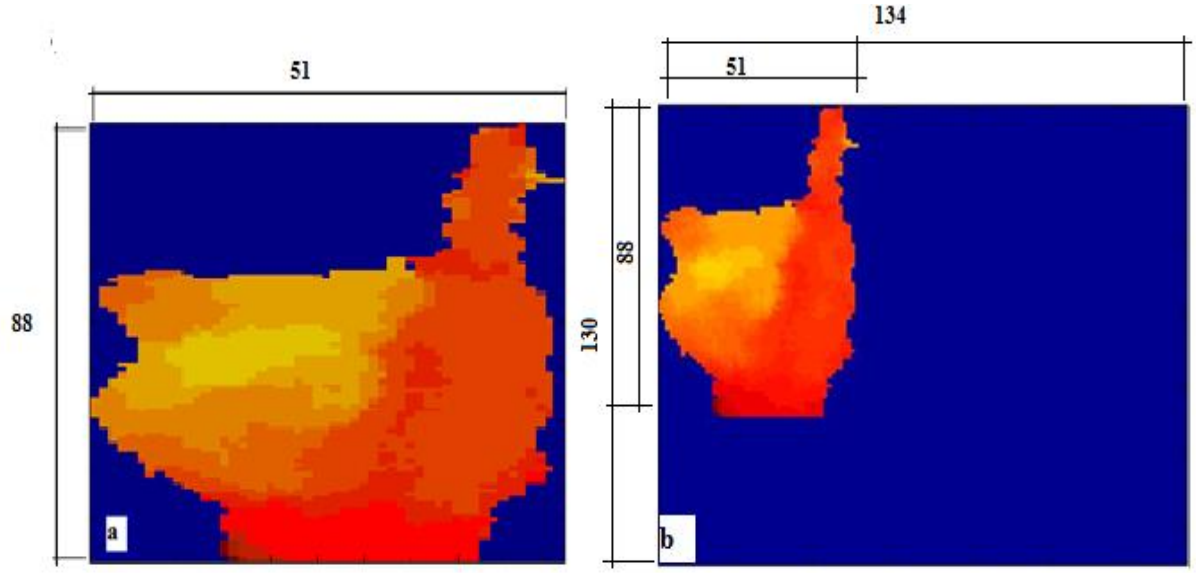


Figura 31- Exemplo de Padronização de tamanho. Indivíduo P1, configuração CM5. (a) imagem original com tamanho 85x51, (b) imagem padronizada com tamanho 130x134.

4.3.5 Normalização

A normalização constitui-se na última etapa do pós-processamento. A operação de normalização consiste em subtrair de todos os *pixels* do objeto o valor da menor profundidade diferente de zero presente na imagem (Figura 32). Com esse procedimento, as imagens de uma mesma CM obtidas a diferentes distâncias do Kinect® passam a ter a mesma referência.

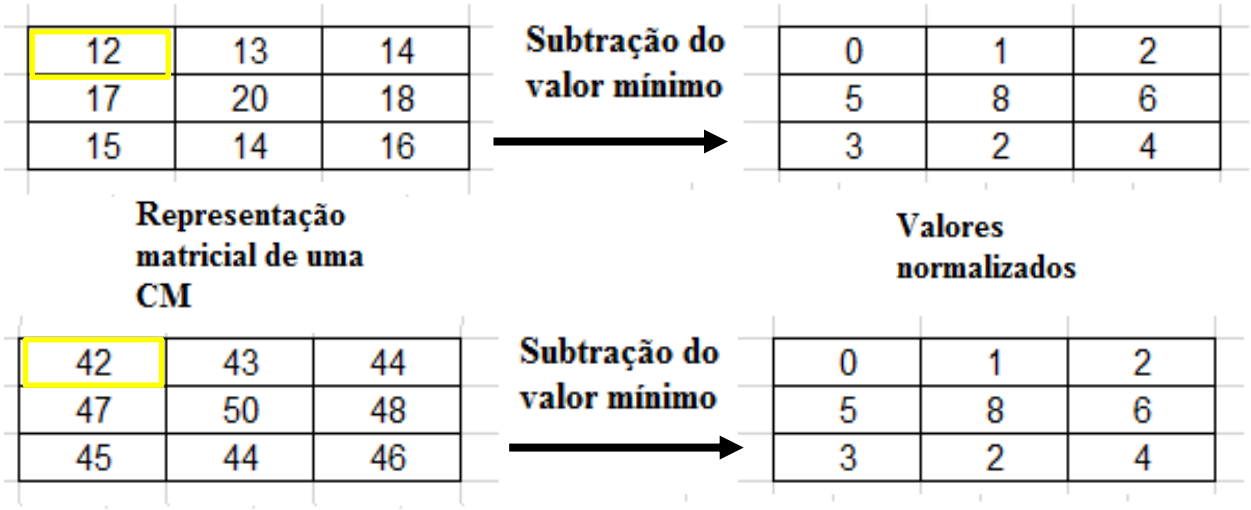


Figura 32-Exemplo de normalização. À esquerda representação matricial de um mesmo gesto com mesma faixa dinâmica, mas com valores de pixels diferentes; à direita, imagens normalizadas, com valores de pixels semelhantes.

4.4 EXTRAÇÃO DE CARACTERÍSTICAS

Para extração de características das imagens utilizou-se a técnica 2D2LDA vista na seção 3.4.2. Por meio dessa técnica consegue-se a redução de tamanho de uma imagem bidimensional em ambas as dimensões.

Na implementação da técnica, o conjunto original das imagens foi dividido em dois conjuntos de igual tamanho, sendo um de treinamento e outro de teste. As matrizes de projeções descritas na seção 3.4.2 são construídas utilizando-se apenas as imagens de treinamento. As dimensões utilizadas para C_k são 5x5, 10x10, 15x15 e 20x20 (Figura 33).

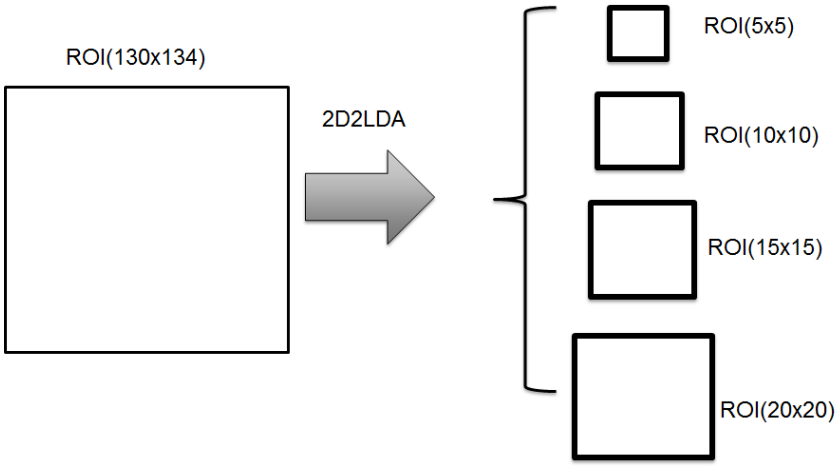


Figura 33- Redução de dimensionalidades usadas.

4.5 CLASSIFICAÇÃO

O classificador utilizado foi o *k*-vizinhos mais próximos (kNN), descrito na seção 3.5.1. A fase de treinamento do algoritmo consiste em construir o conjunto de treinamento, constituído por 61 subconjuntos. Cada subconjunto correspondente a uma CM. Cada subconjunto contém metade das imagens da CM correspondente, ou seja, 100 imagens. As outras 100 imagens formam o conjunto de teste. Dado um novo padrão, o classificador calcula a distância desse novo padrão em relação a cada padrão no conjunto de treinamento, criando

uma lista ordenada, onde o padrão do conjunto de treinamento mais próximo do novo padrão encontra-se no topo da lista e o padrão mais longe, na base da lista (THEODORIDIS e KOUTROUMBAS, 2008). Para o valor de *k=1*, a classe a que pertence o padrão corresponde a classe do padrão situado no topo da lista. Quando $k > 1$, o novo padrão será classificado como sendo da classe que mais aparecer entre os k primeiros elementos da lista. Foram realizados vários experimentos, variando-se o valor de *k* de 1 a 10.

As métricas de distância usadas no classificador foram a distância Euclidiana (equação (26)) e a distância *Manhattan* (equação (27)). Nessas equações, D_{ij} representa a distância entre os padrões i e j, n representa a dimensão do vetor de características.

A Figura 34 representa um diagrama de blocos do classificador. Nesse diagrama 6100 amostras são utilizadas no conjunto de teste, enquanto que 6100 amostras são utilizadas no conjunto de treinamento. Nesse diagrama, denomina-se o bloco indutor como um bloco que calcula a distância (Euclidiana ou *Manhattan*) de uma imagem do conjunto de treinamento em relação as imagens do conjunto de teste.

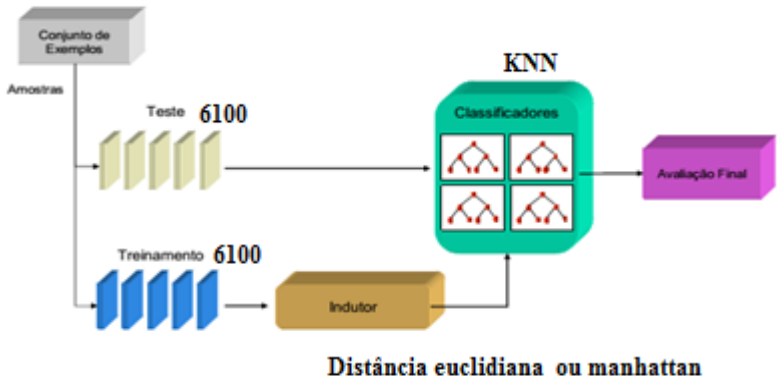


Figura 34- Diagrama da metodologia de validação do classificador.

5 RESULTADOS

No âmbito de trabalhos voltados para reconhecimento da Língua de Sinais, o presente trabalho automatiza o reconhecimento das 61 configurações de mãos que é um dos parâmetros da LIBRAS na formação dos gestos.

Neste capítulo serão apresentados os principais resultados. A seção 5.1, apresenta resultados experimentais das etapas de segmentação e pós-processamento nas imagens de treinamento e teste. A seção 5.2 apresenta resultados da taxa de acerto média relacionados com tamanho do vetor de características assim como o valor de k vizinhos mais próximos.

5.1 RESULTADOS NA SEGMENTAÇÃO E PÓS-PROCESSAMENTO

Nesta seção são mostrados resultados relativos à segmentação (seção 4.2.1) e ao algoritmo de pós- processamento.

5.1.1 Segmentação

As Figura 35, 36 e 37 ilustram exemplos de segmentação para *K=3, K=4 e K=5*, respectivamente, para diferentes configurações, CM_j ($1 \leq j \leq 61$), e indivíduos, P_i ($1 \leq i \leq 10$).

Dois problemas podem ser observados nas Figura 35, 35 e 36: ruídos e erros de segmentação.

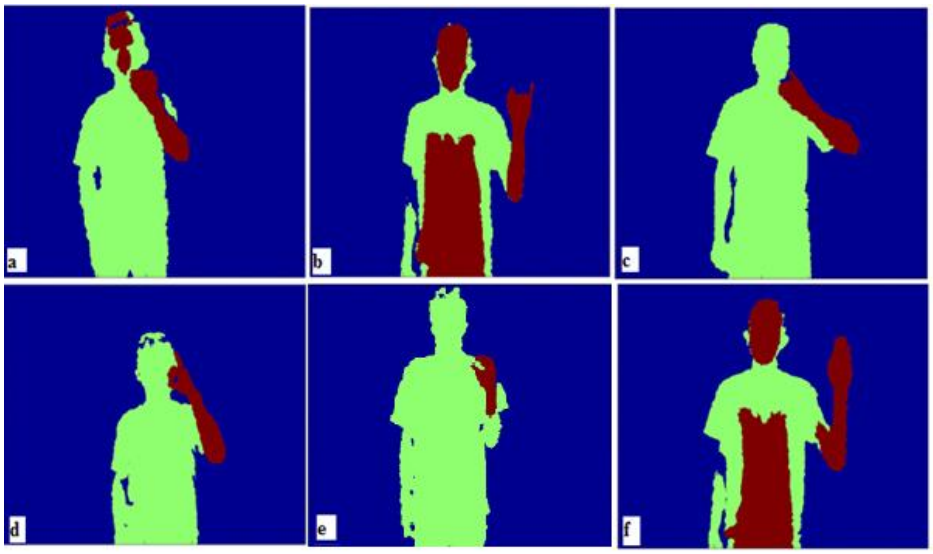


Figura 35- Resultado K-*means* para K=3. (a)P_2,CM_1, (b)P_1, CM_3, (c) P_5, CM_6, (d) P_2, CM_{16}, (e) P_3, CM_{28}, (f) P_1, CM_{26}.

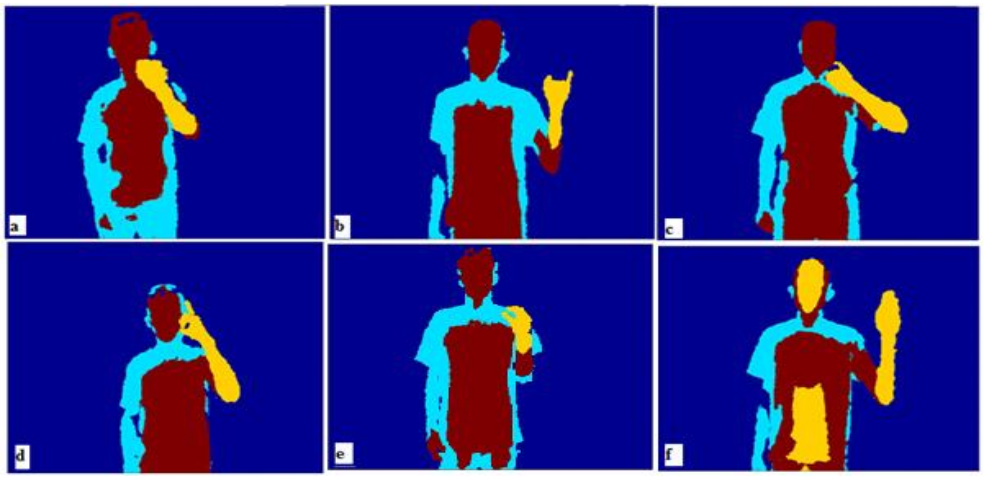


Figura 36 Segmentação K-*means* com K=4. (a)P_2,CM_1, (b)P_1, CM_3, (c) P_5, CM_6, (d) P_2, CM_{16}, (e) P_3, CM_{28}, (f) P_1, CM_{26}.

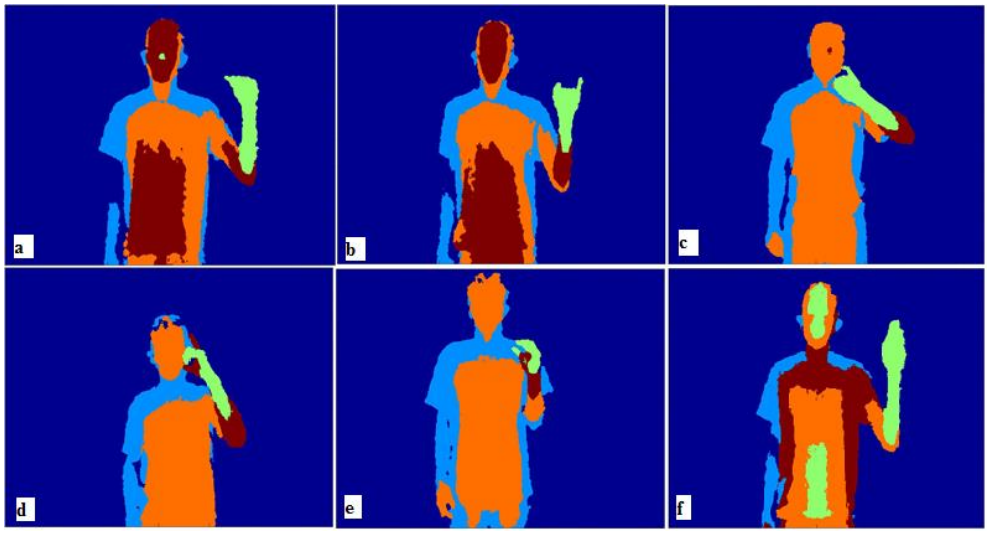


Figura 37- Segmentação K-*means* com K=5:(a) P_2,CM_1, (b)P_1, CM_3 (c) P_5, CM_6, (d) P_2, CM_{16}, (e) P_3, CM_{28}, (f) P_1, CM_{26}.

Observa-se nas Figura 35(a), 35(b), 35(f), 36(f), 37(a) e 37(f) que, quando a mão está no mesmo plano que o corpo, os resultados apresentam ruídos: região da cabeça, barriga e cotovelo são agrupadas em um mesmo grupo. De uma forma geral, com *K=3*, 66,5% do conjunto de imagens apresentaram algum tipo de ruído, enquanto que com *K=4* e *K=5* uma porcentagem de 39,2% e 17,7% das imagens, respectivamente, apresentaram ruídos. O procedimento para filtragem dos ruídos já foi descrito anteriormente no capítulo da metodologia. Ressalta-se que para se encontrar imagens com ruído é feito uma rotina no Matlab® de tal forma que uma imagem é dita ruidosa quando o programa detectar mais de um objeto conectado com vizinhos de 4.

As Figura 37 (d) e 37(e) mostram que, para *K=5*, parte da mão foi rotulada em outro grupo (cor marrom). De uma forma geral, com *K=5* foram encontrados erros de segmentação da mão em 2,12% das imagens, enquanto que, para *K=3 e K=4*, esses erros não foram encontrados.

Baseado nos resultados das Figura 35(d), 36(d) e 37(b), observa-se que, quando a mão do usuário está bem à frente do seu corpo, resultados satisfatórios da etapa de segmentação são

obtidos, ou seja, o antebraço e a mão são segmentados em um único grupo, de forma separada das outras regiões do corpo.

Com base nessas análises, conclui-se que para *K=3* a segmentação da ROI apresenta muitos ruídos, para *K=5* a segmentação foi errônea em algumas imagens. Portanto, a melhor segmentação foi obtida utilizando-se *K=4*.

Os resultados a serem obtidos para o classificador são baseadas nas imagens resultantes da segmentação com *K=4.*

5.2 RESULTADOS DA CLASSIFICAÇÃO

As medidas utilizadas para avaliação da metodologia proposta para reconhecimento dos gestos são: taxa de acerto no reconhecimento de cada CM, taxa média de acerto considerando as taxas de acerto de todas as CM, com o respectivo desvio padrão.

As Figura 38 e 39 ilustram as variações da taxa média de acerto em função dos parâmetros utilizados na etapa de classificação: tamanho da matriz de características, C_k, do algoritmo 2D2LDA e valor de *k* no algoritmo kNN. Cada curva corresponde a um tamanho diferente para a matriz de características. Na Figura 38 é utilizada a distância Euclidiana no algoritmo kNN, enquanto que na Figura 39 é utilizada a distância *Manhattan* no algoritmo kNN. Importante notar que o k se refere ao número de vizinhos mais próximos e o K se refere ao número de *clusters.*

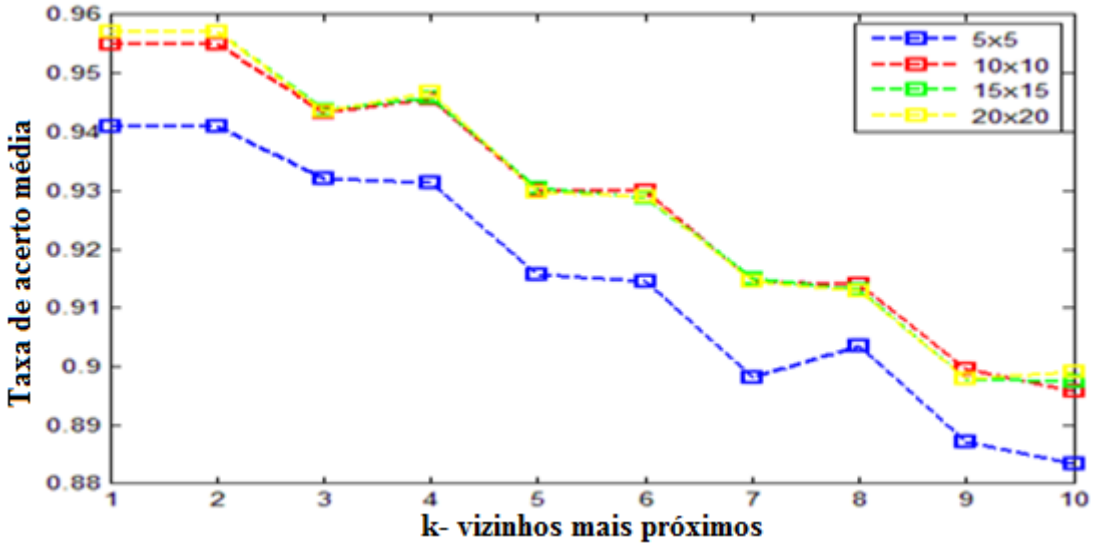


Figura 38-Curvas de taxa média de acerto em função do tamanho da matriz de características do método 2D2LDA e do número de vizinhos mais próximos, k, do método KNN, utilizando a distância euclidiana nesse último método.

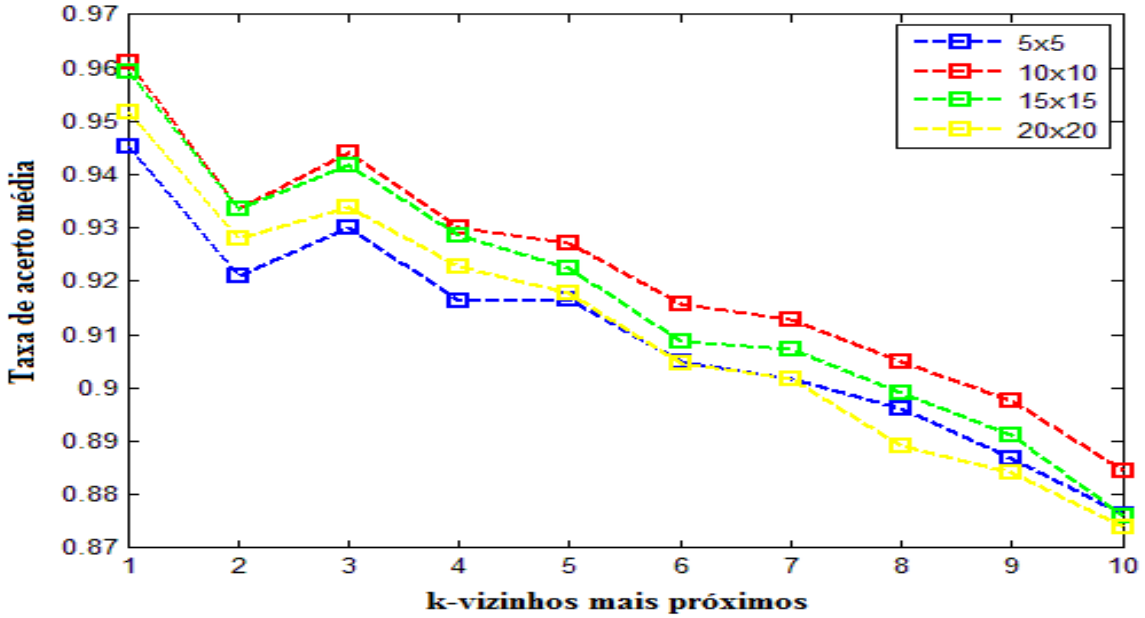


Figura 39- Curvas de taxa média de acerto em função do tamanho da matriz de características do método 2D2LDA e do número de vizinhos mais próximos, k, do método KNN, utilizando a distância *manhattan* nesse último método.

A partir da Figura 39 observa-se que: o maior valor para a taxa média de acerto é obtida para 1NN- *Manhattan* e corresponde ao valor de 96,10%, com C_k de dimensão 10x10.

Os resultados obtidos para a taxa média de acerto em função do tamanho da matriz de características do método 2D2LDA e do número de vizinhos mais próximos, k, do método

kNN, e dos desvios padrões correspondentes são mostrados nas Tabelas numeradas de 1 a 4 na sequência.

A Tabela 1 mostra em negrito o melhor resultado usando um vetor de características concatenado da matriz de características de 5x5. Para k=1 a média de acerto é de 94,52% com desvio padrão de 3,23.

Tabela 1- Resultados da taxa média de acerto para matriz 5x5.

nº de vizinhos (k)	Medida de distância			
	Euclidiana		*Manhattan*	
	Média de acerto	Desvio padrão	Média de acerto	Desvio padrão
1	94,09836	3,59282	**94,52459**	3,232045
5	90,06557	5,159327	91,65574	4,633751
10	86,16393	6,650971	87,63934	6,169818
15	81,85246	8,097431	84,19672	6,991376

A Tabela 2 realça em negrito o melhor resultado usando um vetor de características concatenado da matriz de características de 10x10. Para k=1 a média de acerto é de 96,10% com desvio padrão de 2,63.

Tabela 2-Matriz da taxa média de acerto de características 10x10.

nº de vizinhos (k)	Medida de distância			
	Euclidiana		*Manhattan*	
	Média de acerto	Desvio padrão	Média de acerto	Desvio padrão
1	95,5082	2,956643	**96,09836**	2,634605
5	91,72131	4,680378	92,70492	4,350784
10	87,22951	6,125447	88,45902	6,157958
15	83,08197	7,486151	84,95082	7,282195

Tabela 3 realça em negrito o melhor resultado usando um vetor de características concatenado da matriz de características de 15x15. Para k=1 a média de acerto é de 95,93% com desvio padrão de 2,77.

Tabela 3- Resultado da taxa média de acerto para matriz 15x15.

nº de vizinhos (k)	Medida de distância			
	Euclidiana		*Manhattan*	
	Média de acerto	Desvio padrão	Média de acerto	Desvio padrão
1	95,70492	2,955097	**95,93443**	2,774999
5	91,7377	4,797204	92,2623	4,555313
10	87,42623	6,036661	87,59016	6,891137
15	83,39344	7,740205	84,14754	8,073101

A Tabela 4 realça em negrito o melhor resultado usando um vetor de características concatenado da matriz de características de 20x20. Para k=1 a média de acerto é de 95,70% com desvio padrão de 2,99.

Tabela 4- Resultado da taxa média de acerto para matriz 20x20.

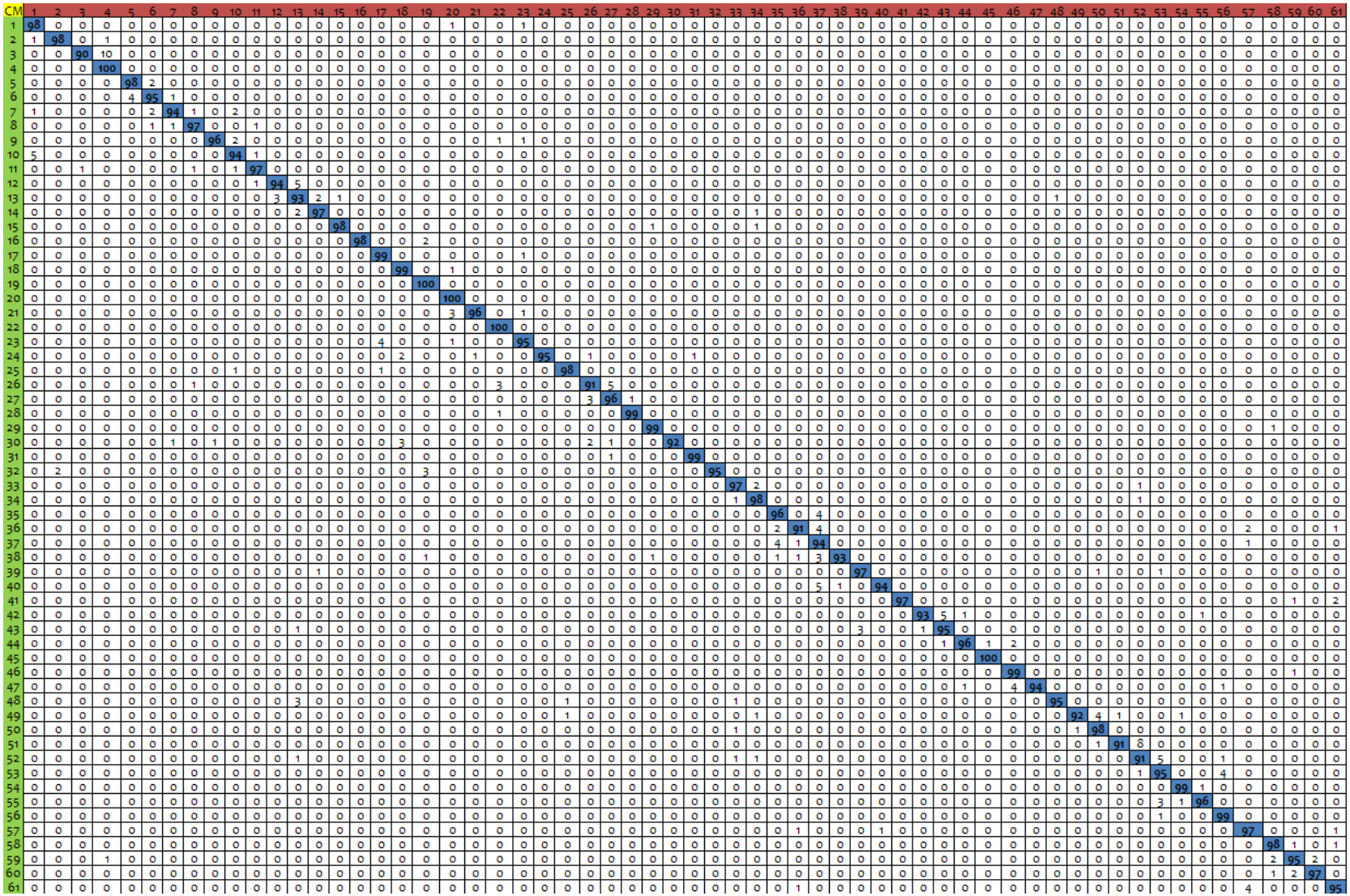
nº de vizinhos (k)	Medida de distância			
	Euclidiana		*Manhattan*	
	Média de acerto	Desvio padrão	Média de acerto	Desvio padrão
1	**95,70492**	2,993678	95,16393	3,364133
5	91,98361	4,740795	91,78689	4,865788
10	87,47541	6,173998	87,39344	6,876614
15	83,40984	7,918389	83,54098	8,213491

A matriz de característica cuja dimensão é 10x10 apresenta melhores taxas de acerto levando em consideração todos os valores de k quando usado a distância *Manhattan*. Esse tamanho maximiza as distâncias entre as classes e diminui a distância entres os padrões na classe.

A Figura 40 mostra, para k=1 e para a matriz de características 10x10, a matriz de confusão para a classificação de cada CM. Na diagonal principal da tabela estão os totais de acerto. Cada elemento (i, j) (linha i, coluna j) corresponde ao número de CM’s rotulados como i que foram classificados como j.

Por exemplo, a somatória da nona linha expressa que existem 100 gestos da CM 9, sendo que 96 foram classificados corretamente, dois foram classificados como CM 10, um foi classificado como CM 22 e um foi classificado como CM 23. Analisando a Figura 40 observa-se:

- As CM 4, 19, 20, 22, 45 foram as que apresentaram taxas de acerto máximas: 100 acertos.
- A CM 3 foi a que apresentou o menor número de acertos: 90 acertos, sendo registrados 10 falsos positivos para CM 4.
- A CM 51 correspondeu ao maior número de falsos positivo, 8 CM 52 foram classificados como CM 51.



CM	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26	27	28	29	30	31	32	33	34	35	36	37	38	39	40	41	42	43	44	45	46	47	48	49	50	51	52	53	54	55	56	57	58	59	60	61
1	98	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
2	1	98	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
3	0	0	90	10	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
4	0	0	0	100	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
5	0	0	0	0	98	2	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
6	0	0	0	0	4	95	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
7	1	0	0	0	0	2	94	1	0	2	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
8	0	0	0	0	0	1	1	97	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
9	0	0	0	0	0	0	0	0	96	2	0	0	0	0	0	0	0	0	0	0	0	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
10	5	0	0	0	0	0	0	0	0	94	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
11	0	0	1	0	0	0	0	1	0	1	97	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
12	0	0	0	0	0	0	0	0	0	0	1	94	5	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
13	0	0	0	0	0	0	0	0	0	0	0	3	93	2	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
14	0	0	0	0	0	0	0	0	0	0	0	0	2	97	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
15	0	0	0	0	0	0	0	0	0	0	0	0	0	0	98	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
16	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	98	0	0	2	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
17	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	99	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
18	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	99	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
19	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	100	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
20	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	100	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
21	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	3	96	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
22	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	100	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
23	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	4	0	0	1	0	0	95	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
24	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	2	0	0	1	0	0	95	0	1	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
25	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	1	0	0	0	0	0	0	0	98	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
26	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	3	0	0	0	91	5	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
27	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	3	96	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
28	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	99	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
29	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	99	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0
30	0	0	0	0	0	0	1	0	1	0	0	0	0	0	0	0	0	3	0	0	0	0	0	0	0	0	2	1	0	0	92	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
31	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	99	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
32	0	2	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	3	0	0	0	0	0	0	0	0	0	0	0	0	0	95	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
33	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	97	2	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0
34	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	98	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0
35	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	96	0	4	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
36	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	2	91	4	0	0	0	0	0	0	0	0	0	0	0	0	0	0	2	0	0	0	1				
37	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	4	1	94	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0				
38	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	1	1	3	93	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0				
39	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	97	0	0	0	0	0	0	0	0	1	0	0	1	0	0	0	0				
40	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	5	1	0	94	0	0	0	0	0	0	0	0	0	0	0	0	0				
41	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	97	0	0	0	0	0	0	0	0	0	1	0	2							
42	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	93	5	1	0	0	0	0	0	0	0	1	0	0	0	0				
43	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	3	0	0	1	95	0	0	0	0	0	0	0	0	0				
44	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	96	1	2	0	0	0	0	0	0	0	0	0				
45	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	100	0	0	0	0	0	0	0	0	0	0	0				
46	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	99	0	0	0	0	0	0	0	0	1	0	0				
47	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	4	94	0	0	0	0	0	1	0	0	0	0				
48	0	0	0	0	0	0	0	0	0	0	0	3	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	95	0	0	0	0	0	0	0	0	0				
49	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	92	4	1	0	0	1	0	0	0				
50	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	98	0	0	0	0	0	0	0	0	0				
51	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	91	8	0	0	0	0	0	0	0				
52	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	91	5	0	0	1	0	0	0				
53	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	95	0	0	4	0	0	0	0	0
54	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	99	1	0	0	0	0	0	0
55	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	3	1	96	0	0	0	0	0	0
56	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	99	0	0	0	0	0
57	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	1	0	0	0	0	0	0	0	0	0	0	97	0	0	0	0	1
58	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	98	1	0	1	
59	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	2	95	2	0		
60	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	2	97	0		
61	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	4	0	0	0	95				

Figura 40- Matriz de confusão (10x10) para 1NN *Manhattan*. As linhas e colunas representam as configurações de mão da LIBRAS.

A Figura 41 mostra valores da taxa de acerto para cada CM e o quão distante as taxas de acerto de cada configuração estão da taxa de acerto média obtida para $C_k = 10$, com k=1 e utilizando a distância *Manhattan*..

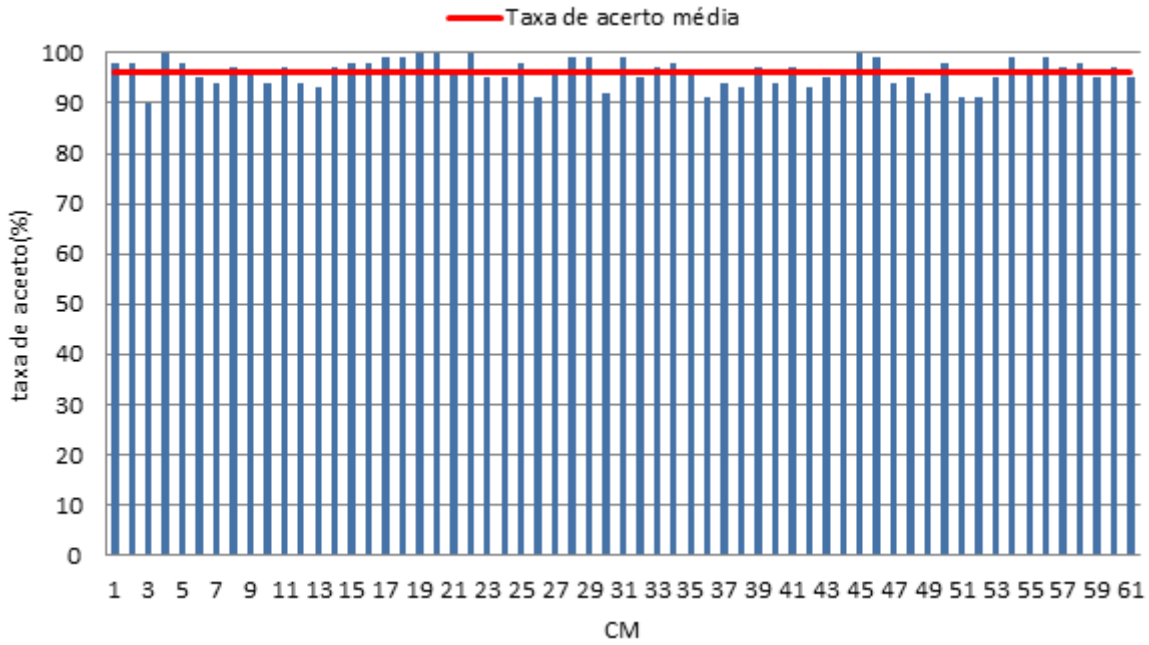


Figura 41- Taxa de acerto para $C_k = 10$, k=1 e utilizando a distância *Manhattan*.

5.3 DISCUSSÃO DOS RESULTADOS

Nas subseções anteriores foi apresentado um rol de experimentos realizados para avaliar a metodologia implementada. Os cenários de teste foram construídos visando a obtenção de resultados que atestam a acurácia e a eficiência do método com respeito aos diferentes elementos de implementação da proposta. Quanto à metodologia empregada, apresenta-se os seguintes comentários:

- A metodologia proposta e implementada redundou num sistema confiável de reconhecimento das configurações de LIBRAS. Esta credibilidade pode ser avaliada pela alta taxa média de acerto obtida.
- A fase de segmentação dos gestos, devido a aplicação da estratégia de segmentação por técnica de agrupamento e posterior aplicação da técnica

transformada de distância Euclidiana para remoção do antebraço, é custosa do ponto de vista de tempo computacional. A técnica de agrupamento, no entanto, é mais apropriada do que a técnica de limiar de profundidade, pois, como já afirmado, a posição da mão do usuário varia bastante, podendo está mais junto ao tronco do indivíduo, o que acarreta dificuldade para se encontrar um limiar que separe a mão do resto da imagem sem o uso de acessórios como luvas e/ou pulseiras.

- A técnica de remoção de antebraço usada (seção 4.3.3) demostrou ser eficaz para a maioria das CM, porém, na remoção do antebraço referente a CM 51 observou-se que numa pequena quantidade de casos, 10% das imagens, houve um corte excessivo do antebraço. Tal fato decorre do deslocamento do centro da palma da mão para cima, o que interfere no ponto de corte da segmentação do antebraço conforme mostrado na equação (31). No entanto, como mostrado na Figura 42, esse corte excessivo pouco afeta a classificação, pois a região mais discriminativa em uma CM não está no encontro da mão com o antebraço, mas sim na parte superior do gesto.

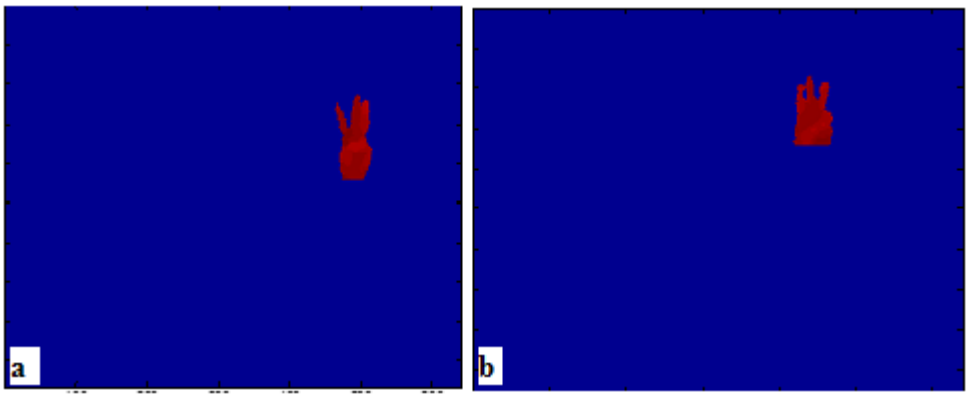


Figura 42- CM 51 com corte excessivo do antebraço. (a) ilustra um gesto sem excesso de corte. (b) ilustra gesto com corte excessivo.

- Os erros de reconhecimento mais frequentes ocorreram no reconhecimento dos seguintes pares: CM 36 e CM 37; CM 51 e CM 52. Na Figura 43 mostram-se as semelhanças das CMs 51 e 52 em uma imagem de profundidade do Kinect®.

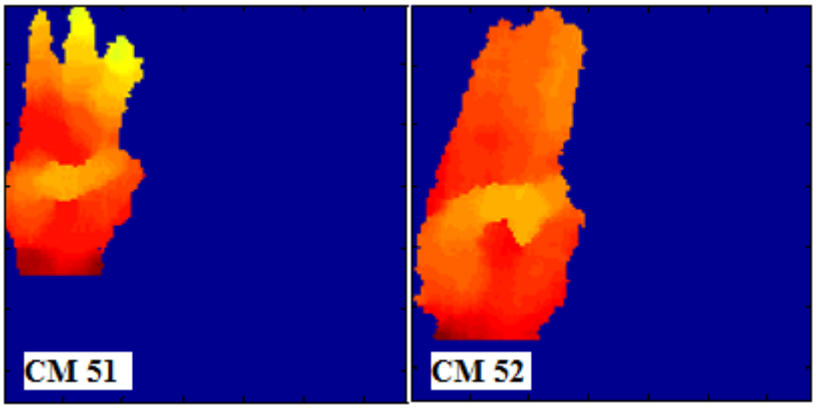


Figura 43- Ilustra a similaridade entre as CMs 51 e 52.

- A CM 3 teve baixa taxa de acerto. Atribui-se tal ocorrência ao posicionamento equivocado da mão realizada pelo usuário $P2$ no processo de construção da base de dados. Tal erro de posicionamento fez com que a CM 3 se parecesse com a CM 4.

As principais conclusões desse trabalho são listadas na sequência.

6 CONCLUSÕES E TRABALHOS FUTUROS

A Língua de sinais não é só uma forma natural de comunicação entre surdos e deficientes auditivos, é também um importante mecanismo de inclusão que deve ser conhecido e estudado pela sociedade como um todo. Semelhantemente às línguas orais, a representação por sinais constitui-se em uma complexa estrutura linguística, fornecendo recursos expressivos suficientes que permitam aos seus usuários expor ideias em relação a quaisquer assuntos e situações.

O estudo apresenta uma metodologia para reconhecimento de configurações de mãos utilizadas na LIBRAS, baseado em imagens de profundidade. A principal diferença entre este trabalho e outros já apresentados são os seguintes: uma base robusta com 12200 imagens da LIBRAS foi gerada para treinamento e teste; fatores de luminosidade não limitam a aplicação desenvolvida. Vetores de características com diferentes tamanhos (400, 225, 100 e 25) são gerados utilizando-se a técnica 2D2LDA. Vetores contendo 100 e 225 características foram os que obtiveram os melhores resultados de classificação com o classificador kNN. Para esse classificador, os melhores resultados na classificação foram obtidos com *k* (número de vizinhos) igual a 1 e usando a distância *Manhattan* como medida de similaridade. Os resultados mostram que a taxa média de acerto é de 96,10%, com um desvio padrão de 2,63, para vetores com dimensão 100 e 95,93% com um desvio padrão de 2,77, para uns vetores com dimensão 225. Os valores desses desvios padrões podem ser explicados, principalmente, pelas configurações 36 e 51, que, devido à baixa resolução do *Kinect*®, acabam gerando confusão com CM 37 e 52, respectivamente, acarretando erros no algoritmo de classificação.

Os outros trabalhos já publicados na literatura (MARAQA *et al*. (2012), ZHOU *et al*. (2013)) fizeram uso de características diferentes, como os Momentos invariantes de Hu, Orientação de Histograma Local e Descritores de Fourier que focaram em objetivos diferentes,

como o reconhecimento de sinais ou sentenças e não configurações de mão. Desta forma, não faz sentido a comparação dos resultados obtidos no atual estudo com os resultados desses trabalhos previamente publicados.

Por fim, registra-se a robustez do trabalho desenvolvido, haja vista a imensa base de dados construída e utilizada pelo grupo do PPGEE que trabalha com o reconhecimento de gestos de língua de sinais.

Como continuidade desse trabalho propõe-se:

- A utilização de outros classificadores como redes convolucionais, uma ferramenta de aprendizado profundo, criada especialmente com a finalidade de reconhecimento de padrões em imagens;
- O reconhecimento dos outros fonemas da LIBRAS;
- O reconhecimento completo de mensagens em LIBRAS.

REFERÊNCIAS

BORGEFORS, G. Distance transformations in digital images. **Computer Vision, Graphics, and Image Processing,** v. 34, n. 3, p. 344-371, 6// 1986.

BRAGATTO, T. A. C.; RUAS, G. I. S.; LAMAR, M. V. Real-time video based finger spelling recognition system using low computational complexity Artificial Neural Networks. In: **Telecommunications Symposium, 2006 International**. IEEE, 2006. p. 393-397.

CARNEIRO, Alex Torquato Souza. **Sistema de Reconhecimento do Alfabeto da LIBRAS por Visão Computacional e Redes Neurais.** 2010. 98 f., il. Dissertação (Mestrado em Engenharia de Teleinformática)- Departamento de Engenharia de Teleinformática, Universidade Federal do Ceará, Fortaleza, 2010.

CHAO, S. et al. Latent support vector machine for sign language recognition with Kinect. **Image Processing (ICIP), 2013 20th IEEE International Conference** on, 2013, 15-18 Sept. 2013. p.4190-4194.

DA FONTOURA COSTA, Luciano; CESAR JR, Roberto Marcond. **Shape analysis and classification: theory and practice**. CRC press, 2010.

DEIMEL, B.; SCHRÖTER, S. **Improving Hand Gesture Recognition Via Video Based Methods for the Separation of the Forearm from the Human Hand**. Dekanat Informatik, Univ., 1998.

GOLDFELD, Márcia. A criança surda. **Linguagem e Cognição numa Perspectiva Sociointeracionista**. São Paulo: p/exus, 2003.

GONZALEZ, R. C.; WOODS, R. E. **Digital image processing**: Prentice hall Upper Saddle River, NJ 2002.

GUIMARAES, Cayley et al. Structure of the Brazilian sign language (Libras) for computational tools: citizenship and social inclusion. In: **Organizational, Business, and Technological Aspects of the Knowledge Society**. Springer Berlin Heidelberg, 2010. p. 365-370.

HASSANI, A. Z. et al. Touch versus in-air hand gestures: evaluating the acceptance by seniors of human-robot interaction. In: (Ed.). **Ambient Intelligence**: Springer, 2011. p.309-313.

JAIN, A. K. Data clustering: 50 years beyond K-means. **Pattern Recognition Letters,** v. 31, n. 8, p. 651-666, 2010.

JUNGONG, H. et al. Enhanced Computer Vision With Microsoft Kinect Sensor: A Review. **Cybernetics, IEEE Transactions on,** v. 43, n. 5, p. 1318-1334, 2013.

KUBASKI, C.; MORAES, V. P. O bilinguismo como proposta educacional para crianças surdas. In: **IX Congresso Nacional de Educação–EDUCERE–III Encontro Sul Brasileiro de Psicopedagogia, PUCPR, PR**. 2009. p. 3415.

LI, M.; YUAN, B. 2D-LDA: A statistical linear discriminant analysis for image matrix. **Pattern Recognition Letters,** v. 26, n. 5, p. 527-532, 2005.

LAMAR, Marcus V.; BHUIYAN, Md Shoaib; IWATA, Akira. Hand gesture recognition using morphological principal component analysis and an improved CombNET-II. In: **Systems, Man, and Cybernetics, 1999. IEEE SMC'99 Conference Proceedings. 1999 IEEE International Conference on**. IEEE, 1999. p. 57-62.

MALASSIOTIS, S.; STRINTZIS, M. G. Real-time hand posture recognition using range data. **Image and Vision Computing,** v. 26, n. 7, p. 1027-1037, 2008.

MARAQA, M. et al. Recognition of Arabic Sign Language (ArSL) using recurrent neural networks. **Journal of Intelligent Learning Systems and Applications,** v. 4, p. 41, 2012.

MARCOTTI, Paulo et al. Interface para Reconhecimento da Língua Brasileira de Sinais. In: **Anais do Simpósio Brasileiro de Informática na Educação**. 2007. p. 482-489.

MAUNG, Tin Hninn Hninn. Real-time hand tracking and gesture recognition system using neural networks. **World Academy of Science, Engineering and Technology**, v. 50, p. 466-470, 2009.

HU, Ming-Kuei. Visual pattern recognition by moment invariants. **Information Theory, IRE Transactions on**, v. 8, n. 2, p. 179-187, 1962.

MITRA, Sushmita; ACHARYA, Tinku. Gesture recognition: A survey. **Systems, Man, and Cybernetics, Part C: Applications and Reviews, IEEE Transactions on**, v. 37, n. 3, p. 311-324, 2007.

NOUSHATH, S.; KUMAR, G. Hemantha; SHIVAKUMARA, P. (2D) 2 LDA: An efficient approach for face recognition. **Pattern Recognition**, v. 39, n. 7, p. 1396-1400, 2006.

OTSU, N. A threshold selection method from gray-level histograms. **Automatica,** v. 11, n. 285-296, p. 23-27, 1975.

PEIXOTO, Adelailson; VELHO, Luiz Carlos. **Transformadas de distância**. PUC, 2000.

PIMENTA, N.; DE QUADROS, R. M. Curso LIBRAS 1 4a Ediçao. **Editora Vozes**, 2010.

PORFIRIO, Andres Jesse et al. LIBRAS Sign Language Hand Configuration Recognition Based on 3D Meshes. In: **Systems, Man, and Cybernetics (SMC), 2013 IEEE International Conference on**. IEEE, 2013. p. 1588-1593.

PROKOP, R. J.; REEVES, A. P. A survey of moment-based techniques for unoccluded object representation and recognition. **CVGIP: Graphical Models and Image Processing,** v. 54, n. 5, p. 438-460, 1992.

RAKUN, Erdefi et al. Combining depth image and skeleton data from Kinect for recognizing words in the sign system for Indonesian language (SIBI [Sistem Isyarat Bahasa Indonesia]). In: **Advanced Computer Science and Information Systems (ICACSIS), 2013 International Conference on**. IEEE, 2013. p. 387-392.

RIBEIRO, Herbert Luchetti **Reconhecimento de gestos usando segmetação de imgens dinamicas de mãos baseado no modelo de misturas de Gaussianas e cor de pele**. 2006, 144p. Dissertação (Mestrado em Engenharia Elétrica) – Escola de Engenharia de São Carlos, Universidade de São Paulo, 2006.

SHOTTON, Jamie et al. Real-time human pose recognition in parts from single depth images. **Communications of the ACM**, v. 56, n. 1, p. 116-124, 2013.

SOARES, Maria Apª Leite. **A Educação do Surdo no Brasil**.2. ed. Campinas, SP. Autores Associados , 2005

SOUZA, Robson Silva de. **Reconhecimento das Configurações de mão da Língua Brasileira de Sinais-LIBRAS em imagens de profundidade atráves da Análise de Componentes Principais e do classificador K- Vizinhos mais Próximos.** 2015, 110p. Dissertação (Mestrado em Engenharia Elétrica) – Faculdade de Tecnologia, Universidade Federal do Amazonas, 2015

TANG, M. Recognizing hand gestures with Microsoft's kinect. **Palo Alto: Department of Electrical Engineering of Stanford University:[sn]**, 2011.

THEODORIDIS, S.; KOUTROUMBAS, K. **Pattern Recognition, Fourth Edition**. Academic Press, 2008. 900.

WANG, H. Nearest neighbors by neighborhood counting. **Pattern Analysis and Machine Intelligence, IEEE Transactions on,** v. 28, n. 6, p. 942-953, 2006.

YIN, X.; XIE, M. Finger identification and hand posture recognition for human–robot interaction. **Image and Vision Computing,** v. 25, n. 8, p. 1291-1300, 2007.

YUE, W.; RUOYU, Y. Real-time hand posture recognition based on hand dominant line using kinect. **Multimedia and Expo Workshops (ICMEW), 2013 IEEE International Conference on**. IEEE, 2013. p.1-4.

ZAFRULLA, Zahoor et al. American sign language recognition with the Kinect. In: **Proceedings of the 13th international conference on multimodal interfaces**. ACM, 2011. p. 279-286.

ZHOU, R. et al. Robust Part-Based Hand Gesture Recognition Using Kinect Sensor. **Multimedia, IEEE Transactions on,** v. 15, n. 5, p. 1110-1120, 201

APÊNDICE: TRABALHO PUBLICADO

TÍTULO: "**Reconhecimento das configurações de mão de LIBRAS baseado na análise de discriminante de Fisher bidimensional, utilizando imagens de profundidade**", apresentado no Congresso da Sociedade Brasileira de Computação – CSBC 2015 realizado na cidade de Recife-PE.

Reconhecimento das configurações de mão de LIBRAS baseado na análise de discriminante de Fisher bidimensional, utilizando imagens de profundidade

Jonilson R. Santos, Marly G. F. Costa, Cícero F. F. Costa Filho

Centro de Pesquisa e Desenvolvimento de Tecnologia Eletrônica e da Informação – Universidade Federal do Amazonas (CETELI/UFAM)
CEP 69077-000 – Manaus – AM – Brasil

joniroque@hotmail.com,{mcosta,ccosta}@ufam.edu.br

Abstract. *Deaf people communicate using sign language; however, they are just able to communicate with others who have this same knowledge. This communication is limited to other people with knowledge of sign language that, usually, are other deaf people. There are too many people that interact with deaf people in education, health and leisure areas that do not know about sign language. Then the inclusion of deaf people is seriously affected, because they are unable to make themselves understood. This study presents a methodology for automatic gesture recognition; the gestures represent the settings of hands from LIBRAS (Brazilian Language of Signs). The approach consist of constructing an image database by Kinect® sensor. In such images, we applied 2D²LDA technique to reduce their dimension and create new characteristics for classification step. The system is able to segment the hand image and recognize whole 61 settings of Sign Language. The average achieved hit rate was 95.7%. As the capture device is insensitive to light, background and colors of clothes and skin, there are no restrictions about environment.*

Resumo. *As pessoas surdas comunicam-se com outras pessoas por meio da Língua de Sinais. O fato de existirem muitas pessoas nas principais áreas da sociedade (saúde, educação e lazer) que interagem com os surdos, com pouco ou nenhuma familiaridade com uma língua de sinais afeta sobremaneira a inclusão social dos mesmos. Este artigo apresenta uma metodologia para o reconhecimento automatizado dos gestos que representam as configurações de mãos da Língua Brasileira de Sinais - LIBRAS. A abordagem inicial consistiu da construção de um banco de imagens de profundidade adquiridas com o sensor Kinect®. Nessas imagens, foi aplicado a técnica 2D²LDA para a redução de dimensionalidade do conjunto de características usado para classificação. O sistema é capaz de segmentar a mão e reconhecer as 61 configurações de mão da LIBRAS. A taxa média de acerto alcançada foi de 95,70%. Como o dispositivo de captura é insensível a luminosidade, fundo e cores das roupas e da pele, a aplicação desenvolvida adapta-se sem modificações a qualquer ambiente de captura das configurações.*

1. Introdução

A comunicação entre as pessoas é efetuada de diversas formas: oral, escrita ou gestual. A língua de sinais é uma forma gestual que possibilita as pessoas surdas interagirem com outras pessoas. Assim, em vez de transmitir suas ideias acusticamente, eles fazem uso de sinais. Como diversas pessoas que interagem com os surdos em diferentes áreas da sociedade, como saúde, educação e lazer, não possuem o domínio da referida língua, a inclusão social do surdo é afetada negativamente, pois o mesmo não é capaz de fazer-se entender.

Em línguas se sinais, "fonemas" se referem às suas unidades espaciais que funcionam igualmente aos fonemas das línguas orais. As unidades mínimas distintivas em LIBRAS são: Configuração de Mãos (CM), Ponto de Articulação, Movimento-Orientação e Expressão Facial. A Língua Brasileira de Sinais, conta com 61 possíveis CM, as quais são apresentadas na Figura 1 (Pimenta e De Quadros 2010).

Figura 1- Configurações de mão da LIBRAS [Pimenta e De Quadros, 2010].

Para atender a demanda de comunicação dos surdos com a sociedade, foram criados sistemas automáticos que têm o intuito de reconhecer os gestos da língua de Sinais. A maioria desses sistemas é baseada em câmeras sensíveis a luz visível e, portanto, sensíveis às condições de luminosidade. Diversas técnicas destinadas ao reconhecimento de gestos da língua de sinais em imagens adquiridas por meio de câmeras digitais têm sido reportadas na literatura (Ribeiro e Gonzaga, 2006; Bragatto *et al.*, 2006; Carneiro *et al*., 2009; Maraqa *et al*.,2012). Técnicas que usam outros dispositivos de captura sensíveis a profundidade, como por exemplo o *Kinect*®, são relatadas por Rakun *et al*. (2013), Chao *et al.* (2013) e Porfirio *et al.* (2013). Nas abordagens que empregam câmeras digitais a detecção da mão não é uma tarefa trivial. Carneiro *et al.* (2009) montou um banco de imagens de 6 pessoas num ambiente de fundo e iluminação controlados. Para segmentação da mão aplicaram técnicas de limiar simples, no espaço de cor YCbCr (Luminância, Crominância Azul, Crominância Vermelha). Para segmentação da mão, Ribeiro e Gonzaga (2006) usaram os processos de eliminação do fundo com misturas de Gaussianas e limiar de cor de pele. No pós-processamento utilizaram filtros morfológicos para recuperar as falhas oriundas da segmentação. Em contrapartida, Bragatto *et al.*(2006) e Maraqa *et al.*(2012) usaram luvas coloridas que facilitam sobremaneira a segmentação da mão. Além disso, Bragatto *et al.*(2006), na classificação das configurações da mão - CM, extraiu quatro características de cada dedo, provenientes da análise de componentes principais, descritas em Lamar *et al.* (1999). Numa abordagem que utiliza câmera de profundidade, Rakun *et al.* (2013) utiliza o sensor *Kinect*®. As seguintes características são extraídas pelos autores para classificação dos gestos: centroide, eixos maior e menor, orientação e menor polígono convexo da imagem segmentada. Por sua vez, Chao *et al.* (2013) utilizaram o Histograma das Orientações do

Gradiente (HOG), pose do corpo, forma e movimentação da mão para a classificação de configurações de mão em vídeo. Para classificação das CM utilizou-se Máquinas de Vetores de Suporte Latente. O método desenvolvido por esses autores é capaz de encontrar quadros discriminativos e representativos em cada conjunto de vídeo**.**

Porfirio *et al.*(2013) faz o reconhecimento das 61 CM da Língua Brasileira de Sinais - LIBRAS. Na aquisição das imagens, cinco pessoas pousaram para a câmera realizando uma determinada sequência de movimentos. Selecionando, manualmente, a visão frontal e lateral do gesto constrói-se uma malha 3D de cada CM. Ao final, 610 malhas 3D são reconstruídas. A segmentação da mão foi feita manualmente no *software Gimp*. Na extração de características de cada malha aplicaram-se descritores harmônicos esféricos, que são insensíveis a translação, rotação e escala. Para classificação das configurações são usadas máquinas de vetores de suporte (*Support Vector Machines* - SVM) com k*ernel* de funções de base radial (RBF) e k*ernel l*inear. Os melhores resultados alcançados foram 96,67% para o *kernel* RBF e 96,83% para o *kernel* linear.

Este trabalho apresenta uma contribuição para o reconhecimento automatizado da LIBRAS, centrando inicialmente a sua atenção no reconhecimento de um dos seus "fonemas", a configuração da mão (CM). Em trabalhos futuros pretende-se reconhecer todos os demais "fonemas". Esse reconhecimento é feito utilizando-se imagens de profundidade adquiridas pelo sensor *Kinect*®. Para obtenção das variáveis de entrada do classificador aplicou-se a técnica $2D^2LDA$. Através dessa técnica, reduziu-se as dimensões da imagem original para 5x5 pixels, 10x10 pixels, 15x15 pixels e 20x20 pixels, respectivamente. Como classificador, empregou-se o *k–vizinhos mais próximos* (KNN - *K-Nearest Neighbor*).

2. Materiais e Métodos

Nesse trabalho, conforme mostrado na Figura 2, são implementadas todas as fases de um sistema de reconhecimento de padrões. A implantação desse sistema foi realizada no ambiente de simulação Matlab2013®. A seguir descrevemos cada uma dessas etapas.

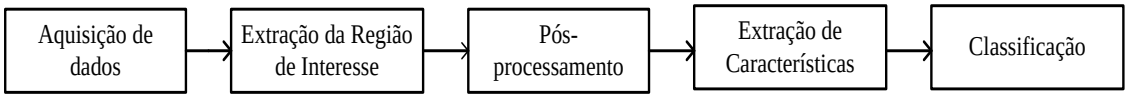


Figura 1- Diagrama em blocos do sistema de visão computacional implementado.

2.1. Aquisição de dados

Nesta etapa é utilizado o *Kinect*® para obter imagens de profundidade, com resolução de 640x480 pixels. A distância da câmera para o usuário é padronizada em 1,40 metros. O banco de imagens é constituído de 12.200 imagens: 10 indivíduos, 61 configurações de mãos, 20 imagens por configuração, por indivíduo. Dentre os 10 indivíduos, dois são mulheres e oito são homens, sendo que sete indivíduos possuem LIBRAS como primeira língua. Procurou-se selecionar indivíduos com mãos de diferentes tamanhos: pequenas médias e grandes. Os usuários foram convidados a rotacioanarem e transladarem as suas mãos lentamente, num intervalo entre 45º e 135º. As 20 imagens por indivíduo e por configuração foram capturadas durante essa movimentação.

2.2. Extração da Região de interesse

2.2.1. Segmentação da mão

Para segmentação da mão foi utilizado a técnica de agrupamento *k-means*, tendo como variável independente a profundidade *z* da imagem. Nessa técnica de segmentação, a eliminação do fundo é importante, pelo fato de impossibilitar a formação de agrupamentos diferentes para cada tipo de fundo. Portanto, inicialmente, utilizou-se a técnica de limiar de Otsu (1975) com o objetivo de subtrair o fundo que, presumidamente, é separado do corpo do indivíduo. O algoritmo de segmentação *k-means* utilizado nesse trabalho pode ser encontrado em Theodoridis e Koutroumbas (2008). Nesse algoritmo, a imagem original I é convertida na forma de vetor. Os principais parâmetros utilizados para esse algoritmo foram: 1) O número de grupos a ser formado, k (3, 4 ou 5); 2) Os centros iniciais, c_1, c_2, c_3, c_4 *e* c_5 , de cada agrupamento, com: $c_1 = 0$, $c_2 =$ *valor mínimo do mapa de profundidade*, $c_3 =$ *valor máximo do mapa de profundidade*, $c_4 =$ *média entre os valores* c_2 e c_3 e $c_5 = c_4/2$. Adotou-se a distância Euclidiana como métrica de distância para formação dos agrupamentos. O grupo referente à região de interesse é selecionado como aquele que tem o centro mais próximo da câmera.

2.3 Pós-processamento

2.3.1. Filtragem

Na maioria das configurações de mão (CM), a mão e o antebraço situam-se à frente do corpo e a segmentação ocorre sem problemas. Entretanto, quando o braço do usuário não está posicionado bem à frente do corpo, e sim lateralmente, a região do antebraço e partes do corpo, como a cabeça, podem ser segmentadas no mesmo grupo. Diz-se, nesses casos, que a segmentação apresenta ruídos. Na Figura 3 apresenta-se um exemplo em que o antebraço é segmentado juntamente com a cabeça. Para filtragem desses ruídos, aplica-se o seguinte procedimento: são medidas as distâncias dos centros dessas regiões segmentadas para o eixo vertical do corpo do usuário. Assumindo que as regiões não estão conectadas entre si, a região contendo o antebraço e a mão (região de interesse - ROI) é aquela que está mais distante do eixo central do usuário. Tal situação é mostrada na Figura 3(a). Na Figura 3(b) apresenta-se o resultado do processo de filtragem.

2.3.2. Rotação e remoção do antebraço

A rotação é utilizada para padronizar todas as ROIs numa mesma direção, a direção vertical. Esse processamento faz-se necessário para a aplicação do algoritmo de extração da mão, conforme será apresentado nas seções seguintes. Para determinação do ângulo de rotação da ROI são utilizados o primeiro e segundo momentos das imagens (Prokop and Reeves, 1992). A matriz de rotação é aplicada na imagem usando a técnica de interpolação do vizinho mais próximo. Na Figura 4(a) apresenta-se um exemplo de ROI rotacionada.

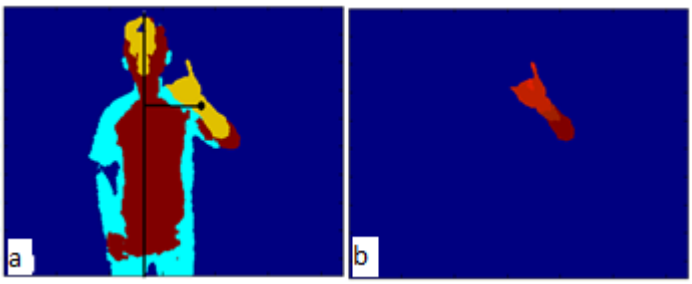


Figura 3- Configuração da mão com antebraço situado lateralmente ao corpo (a) resultado do método *k-means* com 4 grupos mostrando o antebraço segmentado em um mesmo grupo com a cabeça. (b) resultado da filtragem aplicando uma medida de distância.

Após a rotação, a remoção do antebraço é efetuada em duas etapas. O objetivo da primeira etapa é verificar se o cotovelo foi segmentado junto com o antebraço. Nessa etapa os seguintes passos são obedecidos: determina-se o centroide da imagem e traça-se os eixos maior e menor passando por esse centroide (Figura 4(a)); calcula-se a razão entre o eixo maior e o eixo menor. Quando o cotovelo é segmentado junto com o antebraço, a razão entre o eixo maior e o eixo menor é maior do que 3,3. Tal valor foi obtido experimentalmente, a partir da análise do conjunto de imagens do banco. Nos casos dessa razão ser superior ao valor de limiar, elimina-se a parte do antebraço abaixo do centroide (Figura 4(b)) e, posteriormente, aplica-se a transformada de distância. Do contrário, aplica-se simplesmente a transformada de distância.

A segunda etapa objetiva a extração da mão a partir da ROI. Para esse fim, utiliza-se a transformada de distância utilizada por Deimel e Schröter (1998) com o objetivo de determinar o centro e o raio da palma da mão. A transformada de distância consiste em associar a cada pixel p do objeto a menor distância euclidiana, $D(p,q)$, de p para um pixel de borda, q. Gera-se então uma imagem I em que, a intensidade de cada pixel é proporcional a essa distância. As coordenadas do centro, (x_{centro}, y_{centro}) corresponde às coordenadas do pixel de maior intensidade em I. O raio R da palma da mão corresponde ao valor de nível de cinza desse pixel, conforme mostrado na equação (1). A coordena x_{corte} do ponto de corte (x_{corte}, y_{corte}) utilizado para extração da mão (Figura 4(b)) é encontrada multiplicando-se o valor de R por um escalar igual a 1,53 (Deimel e Schröter, 1998) e somando-se com a coordenada x_{centro} da palma da mão, conforme mostrado na equação (2). A região que estiver abaixo do ponto x_{corte} é eliminada da ROI. Na Figura 4(c) mostra-se o resultado da extração da mão aplicando essa transformada.

$$R = \max(I) \tag{1}$$

$$x_{corte} = x_{centro} + 1.53R \tag{2}$$

2.3.3. Padronização

O conjunto de imagens foi padronizado com tamanho 130x134 pixels, que corresponde as maiores dimensões encontradas nas imagens segmentadas. Imagens menores são preenchidas com intensidade zero à direita e abaixo.

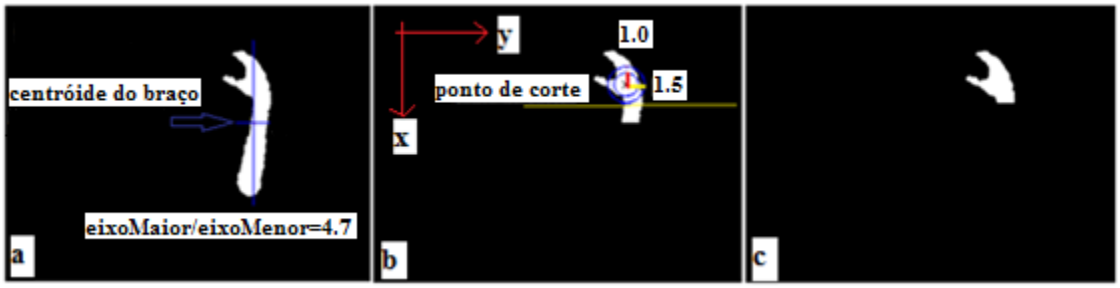


Figura 4 - Seccionamento do braço indivíduo P_1, CM_{29}. (a) cálculo da relação entre os eixos, (b) resultado da primeira da fase, (c) resultado da segunda fase.

2.4. Normalização e Extração de características

Todas as imagens segmentadas e padronizadas são em seguida normalizadas. Para tal, subtrai-se de todos os pixels do objeto o pixel de menor valor diferente de zero. Através desse procedimento, objetiva-se que uma mesma configuração obtida a diferentes distâncias do *Kinect*® correspondam a imagens segmentadas com níveis de cinza semelhantes.

Para extração de características das imagens utilizou-se a técnica 2D^{2}LDA, proposta por Noushath *et al.* (2006). Através dessa técnica consegue-se a redução de tamanho de uma imagem bidimensional em ambas as dimensões. A ideia é projetar as imagens em direções que maximizem um critério de separação entre classes (configurações). Por exemplo, o critério de Fisher (Theodoridis e Koutroumbas, 2008). A seguir descreve-se a operacionalização dessa técnica. Seja uma imagem original I_k, com dimensões mxn. A técnica consiste em se determinar duas matrizes de projeção $X = [x_1, x_2, \dots x_d] e\ Z = [z_1, z_2, .. z_q]$, em que x_i são vetores com dimensões $nx1$ e z_j são vetores com dimensões $mx1$. Tanto os vetores x_i como os vetores z_j são associados a máximos valores de autovalores de matrizes de covariância associados as imagens I_k. Assim, X é uma matriz com dimensão nxd e Z é uma matriz com dimensão mxq. Considerando C_k a matriz com dimensões reduzidas associada a imagem I_k resultante da aplicação de 2D^{2}LDA, a mesma é dada pela equação (3):

$$C_k = Z^T I_k X \tag{3}$$

Os elementos da matriz C_k são a entrada do classificador.

Na implementação da técnica, o conjunto original das imagens foi dividido em dois conjuntos de igual tamanho, sendo um de treinamento e outro de teste. As matrizes de projeções $X\ e\ Z$, descritas anteriormente, são construídas utilizando-se apenas as imagens de treinamento. As dimensões utilizadas para C_k são 5x5, 10x10, 15x15 e 20x20.

2.3. Classificação

O classificador utilizado foi o *k*-vizinhos mais próximos (KNN). A fase de treinamento do algoritmo consiste em construir o conjunto de treinamento, constituído por 61 subconjuntos. Cada subconjunto correspondente a uma CM. Cada subconjunto contém metade das imagens da CM correspondente, ou seja, 100 imagens. As outras 100 imagens formam o conjunto de teste. Dado um novo padrão, o classificador calcula a distância desse novo padrão em relação a cada padrão no conjunto de treinamento, criando uma lista ordenada, onde o padrão do conjunto de treinamento mais próximo do novo padrão

encontra-se no topo da lista e o padrão mais longe, na base da lista (Theodoridis e Koutroumbas, 2008). Para o valor de *k=1*, a classe a que pertence o padrão corresponde a do topo da lista. Então, o novo padrão, será classificado como sendo dessa configuração. Quando $k > 1$, o padrão pertencerá a configuração que mais aparecer entre os *k* primeiros elementos da lista. Foram realizados vários experimentos com os valores de 1 a 10 para o valor de *k*.

3. Resultados

3.1. Segmentação

As Figuras 5, 6 e 7 ilustram exemplos de segmentação para diferentes configurações, CM_j ($1 \leq j \leq 61$), e indivíduos, P_i ($1 \leq i \leq 10$), ao se utilizar 3, 4 e 5 grupos (*k=3, k=4 e k=5*), respectivamente, no algoritmo *k-means*. Observando-se as Figura 5(d), 6(d) e 7(d), verifica-se que, quando a mão do usuário está à frente do corpo, a ROI inclui o antebraço e a mão.

Dois problemas podem ser observados nas figuras 5, 6 e 7: ruídos e erros de segmentação. Os ruídos podem ser vistos nas Figuras 5(a), 5(b), 5(f), 6(f), 7(a) e 7(f). De uma forma geral, com *k=3,* 66,5% das imagens apresentaram algum tipo de ruído, enquanto que com *k=4* e *k=5* uma porcentagem de 39,2% e 17,7% das imagens, respectivamente, apresentaram ruídos. O procedimento para filtragem dos ruídos foi descrito anteriormente. As 7(d) e 7(e) mostram que para *k=5*, parte da mão foi rotulada em outro grupo (cor marrom). De uma forma geral, com *k=5* foram encontrados erros de segmentação da mão em 2,12% das imagens, enquanto que, para *k=3* e *k=4*, esses erros não foram encontrados.

Com base nessas análises conclui-se que para *k=3* a segmentação da ROI apresenta muitos ruídos, para k=5 a segmentação foi errônea em algumas imagens. Portanto, a melhor segmentação foi obtida utilizando-se *k=4*. As análises seguintes são baseadas nas imagens resultantes da segmentação com *k=4*.

3.2. Classificação

A Figura 8 ilustra as variações da taxa média de acerto (taxa obtida considerando a classificação de todas as configurações) do método de classificação em função dos parâmetros utilizados na etapa de classificação: tamanho da matriz de características, C_k, do algoritmo 2D^{2}LDA e valor de *k* no algoritmo KNN.

A partir da Figura 8 observa-se que: o maior valor para a taxa média de acerto, considerando todas as dimensões da matriz C_k, é obtida para 1NN e corresponde ao valor de 95,70% com C_k de dimensão 15x15. A Figura 9 mostra valores da taxa de acerto para cada CM e o quão distante as taxas de acerto de cada configuração estão da taxa de acerto média obtida com C_k com dimensão 15x15 e 1NN. As CM 19, 25, e 56 apresentaram as maiores taxas de acerto. As CM 36 e 51 apresentaram as menores taxas de acerto. Na Tabela 1 mostra-se o valor da máxima taxa média de acerto para cada dimensão da matriz C_k e o desvio padrão observado quando se considera os valores das taxas de acerto em cada configuração. Pode-se observar que: a máxima taxa média de acerto é obtida com a matriz de características C_k com dimensão 15x15 e 20x20; o desvio padrão obtido com a matriz de características C_k com dimensão 15x15 é menor do que o desvio padrão obtido com a matriz de características C_k com dimensão 20x20.

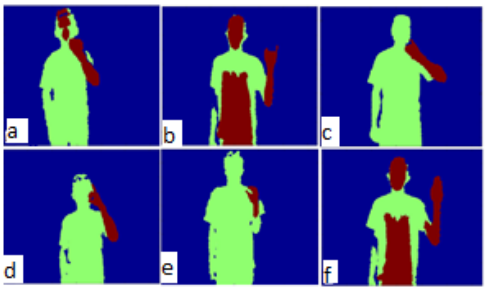


Figura 5 - Segmentação *k-means* com *k=3*. (a)P_2,CM_1, (b)P_1, CM_3, (c) P_5, CM_6, (d) P_2, CM_{16}, (e) P_3, CM_{28}, (f) P_1, CM_{26}.

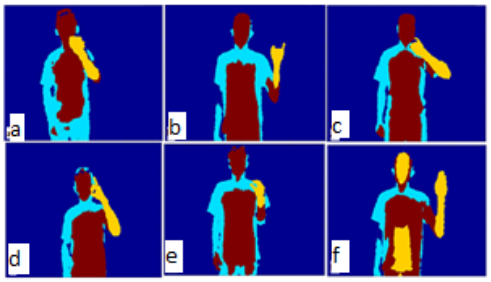


Figura 6 - Segmentação *k-means* com *k=4*:(a)P_2,CM_1, (b)P_1, CM_3 (c) P_5, CM_6, (d) P_2, CM_{16}, (e) P_3, CM_{28}, (f) P_1, CM_{26}.

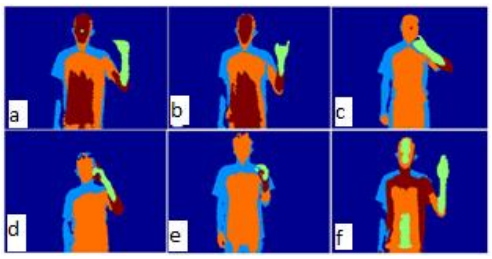


Figura 7 - Segmentação *k-means* com *k=5*:(a)P_2,CM_1,(b)P_1, CM_3, (c) P_5, CM_6, (d)P_2, CM_{16}, (e) P_3, CM_{28}, (f) P_1, CM_{26}.

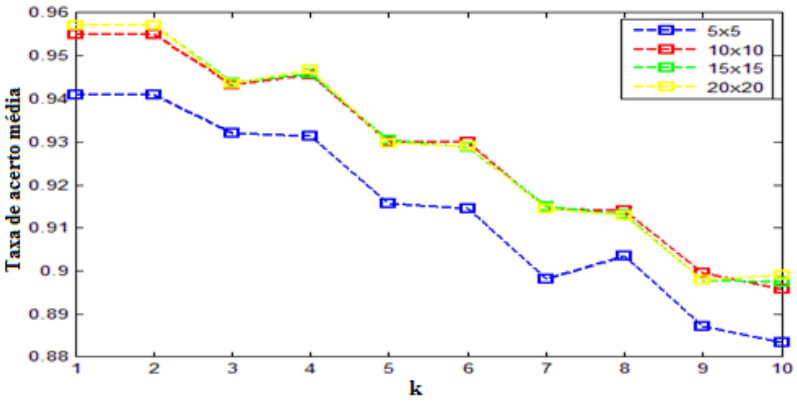


Figura 8 - Taxa de acerto da classificação em função do tamanho da matriz de características, C_K, do algoritmo 2D²LDA e do valor de *k* do método KNN.

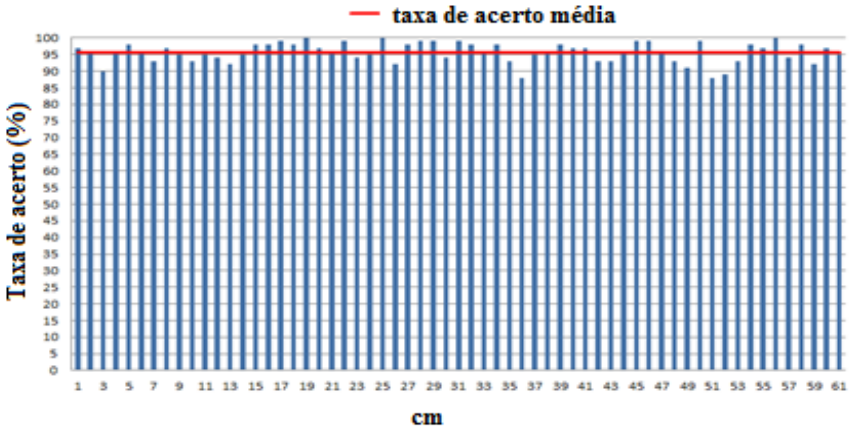


Figura 9- Taxa de acerto para C_k com dimensão 10x10 e 1NN, para cada CM

Tabela 1- Desempenho da classificação para *k=1* (*1NN*).

Matriz Ck	Taxa de acerto média(%)	Desvio padrão
5x5	94,10	3,22
10x10	95,51	2,98
15x15	95,70	2,98
20x20	95,70	3,02

4. Conclusão

O estudo apresenta uma metodologia para reconhecimento de gesto baseado em imagens de profundidade. As principais diferenças entre este trabalho e outros já apresentados são as seguintes: uma base robusta com 12200 imagens da língua LIBRAS foi gerada para treinamento e teste; fatores de luminosidade não limitam a aplicação desenvolvida, uma vez que são utilizadas imagens de profundidade; vetores de características com diferentes tamanhos (400, 225, 100 e 25) são gerados utilizando-se a técnica $2D^2LDA$. Vetores contendo 225 e 400 características foram os que resultaram em melhores resultados de classificação com o classificador KNN. Para esse classificador, os melhores resultados foram obtidos com k igual a 1. Os resultados mostram que a máxima taxa média de acerto é de 95,70%, com um desvio padrão de 2,98 para vetores com dimensão 225, e 95,70% com um desvio padrão de 3,02 para um vetor com tamanho 400. Os valores desses desvios padrões podem ser explicados, principalmente, pelas configurações 36 e 51, que devido à baixa resolução do *Kinect*® acabam gerando confusão com CM 37 e 52, respectivamente, acarretando erros no algoritmo de classificação. O método descrito neste trabalho foi aplicado para o reconhecimento das CM da LIBRAS. Novos esforços serão endereçados futuramente em técnicas que visam reconhecer outros "fonemas" da LIBRAS como a expressão facial.

Agradecimentos

Parte dos resultados apresentados neste trabalho foram obtidos através do Projeto de Pesquisa e formação de recursos humanos, em nível de graduação e pós-graduação, nas áreas de automação industrial, softwares para dispositivos móveis e TV Digital, financiado pela Samsung Eletrônica da Amazônia Ltda., no âmbito da Lei no. 8.387 (art. 2º) /91.

Referências

Bragatto, T. A. C., G. I. S. Ruas, and M. V. Lamar. (2006), Real-time video based finger spelling recognition system using low computational complexity Artificial Neural Networks: Telecommunications Symposium, 2006 International, p. 393-397.

Carneiro, A., P. Cortez, and R. Costa. (2009), Reconhecimento de Gestos da LIBRAS com Classificadores Neurais a partir dos Momentos Invariantes de Hu: Interaction, p. 190-195.

Chao, S., Z. Tianzhu, B. Bing-Kun, X. Changsheng, and M. Tao. (2013), Discriminative Exemplar Coding for Sign Language Recognition With Kinect: Cybernetics, IEEE Transactions on, v. 43, p. 1418-1428.

Deimel, B., and S. Schröter. (1998), Improving Hand Gesture Recognition Via Video Based Methods for the Separation of the Forearm from the Human Hand, Dekanat Informatik, Univ.

Lamar, M. V., M. S. Bhuiyan, and A. Iwata. (1999), Hand gesture recognition using morphological principal component analysis and an improved CombNET-II: Systems, Man, and Cybernetics, 1999. IEEE SMC '99 Conference Proceedings. 1999 IEEE International Conference on, p. 57-62 vol.4.

Maraqa, M., F. Al-Zboun, M. Dhyabat, and R. A. Zitar. (2012), Recognition of Arabic Sign Language (ArSL) using recurrent neural networks: Journal of Intelligent Learning Systems and Applications, v. 4, p. 41.

Noushath, S., G. Hemantha Kumar, and P. Shivakumara. (2006), (2D)2LDA: An efficient approach for face recognition: Pattern Recognition, v. 39, p. 1396-1400.

Otsu, N. (1975), A threshold selection method from gray-level histograms: Automatica, v. 11, p. 23-27.

Pimenta, N., and R. M. de Quadros. (2010), Curso de LIBRAS 1: iniciante, LSB Vídeo.

Porfirio, A. J., K. Lais Wiggers, L. E. S. Oliveira, and D. Weingaertner. (2013), LIBRAS Sign Language Hand Configuration Recognition Based on 3D Meshes: Systems, Man, and Cybernetics (SMC), 2013 IEEE International Conference on, p. 1588-1593.

Prokop, R. J., and A. P. Reeves. (1992), A survey of moment-based techniques for unoccluded object representation and recognition: CVGIP: Graphical Models and Image Processing, v. 54, p. 438-460.

Rakun, E., M. Andriani, I. W. Wiprayoga, K. Danniswara, and A. Tjandra. (2013), Combining depth image and skeleton data from Kinect for recognizing words in the sign system for Indonesian language (SIBI [Sistem Isyarat Bahasa Indonesia]): Advanced Computer Science and Information Systems (ICACSIS), 2013 International Conference on, p. 387-392.

Ribeiro, H. L., and A. Gonzaga. (2006), Reconhecimento de gestos de mão usando o algoritmo GMM e vetor de características de momentos de imagem.

Theodoridis, S., and K. Koutroumbas. (2008), Pattern Recognition, Fourth Edition, Academic Press, 900 p.